Metadatenstandards im Kontext sozialwissenschaftlicher Daten

Jensen, Uwe; Zenk-Möltgen, Wolfgang; Wasner, Catharina

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Jensen, U., Zenk-Möltgen, W., & Wasner, C. (2019). Metadatenstandards im Kontext sozialwissenschaftlicher Daten. In U. Jensen, S. Netscher, & K. Weller (Hrsg.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (S. 151-178). Opladen: Verlag Barbara Budrich. https://doi.org/10.3224/84742233.10

Auszug aus dem Buch:

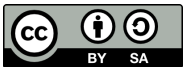
Uwe Jensen
Sebastian Netscher
Katrin Weller (Hrsg.)

Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten

Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten

Verlag Barbara Budrich
Opladen • Berlin • Toronto 2019

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.



Dieses Buch steht im Open-Access-Bereich der Verlagsseite zum kostenlosen Download bereit (https://doi.org/10.3224/84742233).
Eine kostenpflichtige Druckversion (Print on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-8474-2233-4 (Paperback)
eISBN 978-3-8474-1260-1 (eBook)
DOI 10.3224/84742233

Umschlaggestaltung: Bettina Lehfeldt, Kleinmachnow – www.lehfeldtgraphic.de
Lektorat: Nadine Jenke, Potsdam
Satz: Anja Borkam, Jena – kontakt@lektorat-borkam.de
Titelbildnachweis: Foto: Florian Losch
Druck: paper & tinta, Warschau
Printed in Europe

9. Metadatenstandards im Kontext sozialwissenschaftlicher Daten

Uwe Jensen, Wolfgang Zenk-Möltgen und Catharina Wasner

Die transparente und nachvollziehbare Dokumentation von Forschungsdaten und ihres Entstehungskontextes stellt einen wesentlichen Beitrag zu deren Auffindbarkeit, Verständlichkeit, Reproduzierbarkeit und langfristigen Nutzung dar. Werden z.B. in einem Datensatz die Antworten auf eine Frage numerisch codiert (z.B. 1 und 2), müssen Nutzende wissen, was diese Zahlen bedeuten (z.B. Ja und Nein), welcher Inhalt damit verbunden ist (z.B. „Sind Sie wahlberechtigt?“). Datenwerte in einem Datensatz sind nicht selbsterklärend. Sie sind vielmehr der Ausgangspunkt und das Objekt, das systematisch durch Metadaten beschrieben werden muss. Derartige Metadaten – vereinfacht verstanden als *Informationen über Daten* – erstrecken sich von Angaben zu einzelnen Fragen im Fragebogen, deren Antworten als Variablen erfasst werden, bis hin zu den Bedingungen, unter denen die Daten entstanden sind, z.B. durch eine Befragung im Rahmen einer Studie.

Dieses Kapitel behandelt Metadatenstandards und Anwendungsbeispiele, die aufzeigen, welche Arten von Metadaten zur Dokumentation, Zitation sowie zum Auffinden von quantitativen Daten in Katalogen für Forschende in sozialwissenschaftlichen Forschungsprojekten von Bedeutung sind. Aus Sicht des Forschungsdatenmanagements thematisiert dieses Kapitel Metadaten konkret als

> Daten oder Informationen, die in strukturierter Form analoge oder digitale Forschungsdaten (Objekte) dokumentieren. Sie beschreiben, erklären, verorten oder definieren Objekte, Ressourcen und Informationsquellen für die Wissenschaft. Hierdurch helfen sie, Forschungsdaten zu managen, zu erschließen, zu verstehen und zu benutzen (NISO, 2004). (Jensen/Katsanidou/Zenk-Möltgen 2011: 83)

Forschende kommen dabei sowohl als Datenproduzierende als auch als Datennutzende auf unterschiedliche Weise – direkt oder indirekt – mit dem Thema Metadaten in Berührung. Suchen Datennutzende etwa Daten für Sekundäranalysen, greifen sie wahrscheinlich auch auf Datenkataloge zu, um in den dort angebotenen Metadaten zu suchen. Datenproduzierende brauchen wiederum Metadaten, um etwa die Variablen des Datensatzes zur internen Qualitätssicherung und für die Nachnutzung nach Projektende transparent und verständlich zu dokumentieren. Darüber hinaus sind Metadatenstandards für Dateninfrastrukturen (Archive, Repositorien etc.) von großer Bedeutung, um Forschungsdaten systematisch, standardisiert, nachhaltig und miteinander kompatibel zu managen. Um Archivierungssysteme und Datenkataloge in solchen Infrastrukturen sachgerecht zu entwickeln, nutzerfreundlich anzubieten und mit anderen Katalogen zu verbinden, setzen die Anbieter – je nach spezifischen Zweck – unterschiedliche Zusammenstellungen von Metadatenelementen eines Metadatenstandards, sogenannte Metadatenschemata, ein.

Metadatenschemata durchlaufen diverse Entwicklungsphasen, aus denen sich mehr oder weniger verbindliche De-facto- oder Quasi-Standards entwickeln, die auf disziplinspezifischen Praxiserfahrungen und anerkannten Regeln einer wissenschaftlichen Community beruhen. Metadatenstandards können den Status einer Norm erhalten, wie z.B. die *ISO-Norm 15836* des *Dublin Core Metadata Element Set (DCMES 2012)*. Die *Dublin Core Metadata Initiative* (DCMI) entwickelt diesen Standard seit 1994: Waren die Metadaten des Dublin Core Standard zunächst auf die Suche nach Literaturdokumenten und bibliotheksnahe Dienste ausgerichtet, dienen sie heute allgemein der Erschließung von digitalen Objekten im

https://doi.org/10.3224/84742233.10

Internet. Im Laufe der Zeit wurde das Metadatenschema von anderen Disziplinen aufgegriffen, um ihre Objekte, wie z.B. Forschungsdaten, im Web leichter auffindbar zu machen. Allerdings ist es nicht vorgesehen, mit Hilfe des Dublin Core Standard sehr kleinteilige und semantisch reichhaltige Aussagen über Forschungsdaten disziplinspezifisch zu dokumentieren.

Für diese Zwecke entwickelte und etablierte sich ab Mitte der 1990er Jahre das offene Datenmodell der *Data Documentation Initiative* (DDI) als de-facto-Standard zur Dokumentation sozialwissenschaftlicher Forschungsdaten. Ausgehend von den vielfältigen Kontextinformationen, Datenelementen und Datenstrukturen in den Sozialwissenschaften, die durch das Datenmodell definiert werden, ermöglichen entsprechend modellierte DDI-Metadatenstrukturen eine umfassende und differenzierte Beschreibung von Forschungsdaten und deren Entstehungskontext. Metadatenelemente und Metadatenstrukturen und ein zugrunde liegendes Datenmodell stellen, vereinfacht gesagt, den DDI-Metadatenstandard dar, auch kurz DDI Standard genannt.

Das vorliegende Kapitel behandelt sozialwissenschaftlich relevante Metadatenstandards, die Forschungsprojekte darin unterstützen sollen, in Kooperationen mit Dateninfrastrukturen und in Arbeitsteilung mit professionellen Datenmanager/innen ihre Forschungsdaten, Begleitdokumentationen und Projektinformationen

- systematisch, transparent und nachvollziehbar zu dokumentieren,
- zu registrieren und durch Identifikatoren dauerhaft zu zitieren bzw. zu identifizieren,
- in nationalen und internationalen Datenkatalogen und Bibliotheksbeständen zu finden und
- für Replikationen zu sichern bzw. langfristig für die Nachnutzung bereitzustellen.

Die Bearbeitung des Themas erfolgt aus drei Perspektiven, denen jeweils ein Abschnitt in diesem Kapitel gewidmet ist. Abschnitt 9.1 dient der Einführung in das Thema Metadatenstandards und stellt zentrale Begriffe und Konzepte im sozialwissenschaftlichen Kontext vor. Anschließend wird der sozialwissenschaftlich relevante DDI Standard in seinen aktuellen Versionen behandelt. Abschnitt 9.2 erörtert Metadatenstandards zur Auffindbarkeit von elektronischen Ressourcen (Dublin Core Standard), zur Zitation von Datensätzen (DataCite Standard) und zur Dokumentation von sozialwissenschaftlichen Daten auf Studienebene (DDI Standard). Abschließend wird die Frage der Interoperabilität von Metadaten und deren praktischen Nutzen bei Recherchen in Datenkatalogen aus unterschiedlichen Anwendungskontexten erörtert. Diese Diskussion richtet sich primär an Mitarbeitende in Forschung, Lehre und Infrastrukturen, die allgemein an der Nutzung von Metadatenstandards beim Umgang mit sozialwissenschaftlichen Forschungsdaten interessiert sind.

Analog widmet sich Abschnitt 9.3 den verschiedenen Aspekten von Metadaten bei der Dokumentation von Forschungsdaten auf Variablenebene. Dazu wird auf Eigenschaften gängiger Software eingegangen und ihre Fähigkeit zur Erfassung von Metadaten ausgelotet. Außerdem wird beschrieben, wie Metadaten zwischen umfragebasierten Systemen ausgetauscht werden können. Schließlich werden die Möglichkeiten des DDI-Standards zur Variablendokumentation behandelt und die relevanten Module und ihre Metadaten vorgestellt. Abschnitt 9.4 verweist auf ausgewählte Software und Dateiformate zur Verarbeitung und Präsentation von Metadaten. Diese beiden letzten Abschnitte unterstützen den sozialwissenschaftlichen Projektalltag, indem sie über den Umgang mit Metadaten bei der Dokumentation von Variablen und Fragen informieren. Schließlich werden Optionen behandelt, die es Projekten ermöglichen, Metadaten auf Studienebene und auf Variablenebene DDI kompatibel zu erfassen und bereitzustellen.

9.1 Metadaten sozialwissenschaftlicher Studien und Daten

In der langjährigen international anerkannten Praxis sozialwissenschaftlicher Datenarchive werden zwei zentrale Ebenen – die Studienebene und die Variablenebene – zur Dokumentation von Forschungsdaten durch Metadaten unterschieden (vgl. Corti et al. 2014: 38f.).

Auf Studienebene wird mit Hilfe von standardisierten Metadaten, die in einer sogenannten Studienbeschreibung erfasst werden, systematisch der Entstehungskontext der Daten beschrieben, u.a. durch Angaben zum Methodendesign der Studie, Informationen über die beteiligten Forschenden und das Projekt sowie Publikationen zu Datenanalysen. In diesem Kontext werden auch Strukturen und Besonderheiten des Datensatzes dokumentiert, wie z.B. technisches Format der Datenfiles, Anzahl der Fälle und Variablen etc. Diese Metadaten werden für die Recherche in Datenkatalogen genutzt und deshalb häufig auch als *catalog metadata* bezeichnet (vgl. Abschnitt 9.2).

Auf Variablenebene wird jede Variable eines Datensatzes, d.h. die Spalte in der Datenmatrix, und das entsprechende Item der Erhebung durch einen umfangreichen Satz von DDI-Metadaten dokumentiert. Diese sogenannten *rich metadata* werden zur Herstellung von detaillierten Datendokumentationen, etwa in Form von Codebüchern oder Variablenreports, und zur granularen Suche in spezialisierten Recherchesystemen eingesetzt.

Die Bedeutung eines Metadatenstandards liegt in einem einheitlichen Vokabular, das es erlaubt, Informationen über Forschungsdaten verständlich, strukturiert und maschinenverarbeitbar zu erfassen. Bei strukturierten Metadaten handelt es sich im Fall von DDI (und anderen disziplinspezifischen Standards) oft um sehr kleinteilige, granulare – semantisch reichhaltige – Aussagen über die Daten. So umfassen die standardisierten Metadaten einer Variablen Bestandteile wie Variablenformat, -namen, Kodierung erhobener Werte, erklärende Labels etc. sowie ihre Verknüpfungen mit dem Messinstrument, z.B. einem Fragebogen, und deren eigene Metadaten zu Fragetexten, Interviewerанweisungen, Antwortkategorien, Filter usw. Davon abzugrenzen sind semistrukturierte oder unstrukturierte Metadaten, die wichtige Kontextinformationen zur Entstehung der Daten etwa in Form von fließtextbasierten Dokumenten und Materialien beschreiben (z.B. Methodenberichte).

Metadatenstandards legen auch fest, wie Daten und ihre Metadaten (möglichst plattformunabhängig) erstellt, gespeichert und in andere Systeme integriert werden können, um z.B. Datenrecherchen in übergreifenden webbasierten Datenkatalogen zu ermöglichen. Da derartige Systemintegrationen zumeist nicht (vollständig) möglich sind, muss sichergestellt werden, dass Kernmetadaten verschiedener Schemata miteinander kompatibel bzw. interoperabel sind und (möglichst ohne Informationsverlust) aufeinander abgebildet (gemappt) werden können. Kernmetadaten enthalten Informationen, die in einem System vorhanden sein müssen und deshalb das Attribut *verflichtend* (*mandatory)* tragen. Dazu zählen Titel, Art und Produzent einer Ressource (vgl. z.B. Schaukasten 9.3).

Zu berücksichtigen ist, dass Metadaten nicht nur analoge oder digitale Forschungsdaten in strukturierter Form beschreiben, sondern alle möglichen Objekte, wie z.B. Personen, Dokumente, Bücher, Orte, Konzepte oder Webressourcen. Die unterschiedlichen Informationen zu solchen Objekten werden anhand verschiedenster Aspekte kategorisiert und charakterisiert. Dabei lassen sich Metadaten allgemein anhand ihrer Zweckbestimmung in folgende Typen einteilen (Riley 2017; Gilliland 2016; Gartner 2008), die jedoch nicht immer trennscharf sind und zu inhaltlichen Überschneidungen führen können:

- *Beschreibende Metadaten* informieren über diverse Inhalte und formale Eigenschaften, die ein Objekt charakterisieren können, wie z.B. Primärforscher eines Forschungsprojektes, Titel einer empirischen Studie oder die granularen Beschreibungen aller Bestandteile einer Variablen. Beschreibende

Metadaten lassen sich je nach Teilaspekt weiter unterteilen in bibliographische, methodische und literaturbezogene Metadaten usw.

- *Administrative Metadaten* geben Auskunft, wie ein Objekt verwaltet, gefunden und genutzt werden kann. So beschreiben *technische Metadaten* etwa das Dateiformat eines Datenfiles oder die zugrunde liegende Software, mit der Datenanalysen durchgeführt werden. Sogenannte *Preservation Metadaten* beschreiben darüber hinaus unterschiedliche Objekteigenschaften und Ereignisse bei der Sicherung und langfristigen Archivierung von Datenfiles und Kontextinformationen. *Rechtliche Metadaten* informieren schließlich über Nutzungsbedingungen oder andere Verwertungsaspekte der Ressource.
- *Strukturelle Metadaten* beschreiben die Beziehungen zwischen verschiedenen Objekten einer Ressource, z.B. wenn eine Studie aus Datensatz, Fragebogen, Datendokumentation und Methodenbericht besteht.

Metadaten werden in Form von strukturierten Informationen für bestimmte Zwecke und Anwendungskontexte unter Berücksichtigung der disziplinspezifischen Eigenarten von Forschungsdaten systematisch definiert und standardisiert. Ein solcher Metadatenstandard basiert auf der inhaltlichen und technischen Spezifikation eines grundlegenden Metadatenschemas, das oft im XML-Format dargestellt wird. Die Spezifikationen werden von internationalen Konsortien interessierter Organisationen entwickelt und publiziert (vgl. im Linkverzeichnis etwa DDI-C Standard). Die Implementierung eines grundlegenden Metadatenstandards in einem speziellen System wird dann oftmals durch die Publikation des (ggf. modifizierten) praktisch angewendeten Metadatenschemas dokumentiert, wie etwa das Paper *Der GESIS Datenbestandskatalog – und sein Metadatenschema* (Zenk-Möltgen/Habbel 2012) beispielhaft zeigt. Ein solches Schema definiert im Detail für jedes seiner Metadatenelemente (*Terms*) relevante Eigenschaften einer Ressource anhand von zulässigen Angaben (*Values*). Soll etwa eine sozialwissenschaftliche Studie, in deren Kontext Forschungsdaten entstanden sind, in einem Datenkatalog dokumentiert werden, ist u.a. der *Titel der Studie* zu erfassen. Entsprechend definiert etwa ein DDI-Metadatenschema ein Element <title>, das folgende Eigenschaften aufweist:

- Elementnamen: *Titel*,
- Definition des Elements: *Titel der Studie*,
- formale Eigenschaften des Elements: z.B. ist *Pflichtfeld*, ist *sprachabhängig,*
- inhaltlich zulässiger Wert des Elements: z.B. *Eurobarometer 83.4 (2015)*.

Um Forschungsdaten und ihren Entstehungskontext einheitlich und eindeutig zu beschreiben, werden für bestimmte Metadatenelemente standardisierte Terminologien benutzt, um Informationen mit einheitlichen Angaben (*Values*) aufzuführen. Typische Terminologien, die mit DDI-Metadaten auf Studien- oder Variablenebene dokumentiert werden können, zeigen die folgenden Beispiele.

Bei der Planung und Dokumentation von Umfragen wird zur standardisierten Codierung von Sprachen die ISO-Norm 639 eingesetzt, während die ISO 3166 der Codierung von bestehenden Staaten (ISO 3166-1), staatlichen Untereinheiten (ISO 3166-2) und ehemaligen Staaten (ISO 3166-3) dient. Anhand dieser Normen werden u.a. länderspezifische Fragebögen und Sprachversionen komparativer Studien einheitlich dokumentiert. Standardisierte Vokabulare disziplinspezifischer Thesauri und Klassifikationen dienen wiederum der fachspezifischen Indexierung von Literatur, Forschungsdaten und anderen Informationstypen. Die Verschlagwortung sozialwissenschaftlicher Themengebiete erfolgt etwa mit Hilfe des Vokabulars des Thesaurus Sozialwissenschaft (TheSoz), während wirtschaftswissenschaftliche Themen mit dem Standard-Thesaurus Wirtschaft (STW) beschrieben werden können. Die Gemeinsame Normdatei (GND) dient wiederum der einheitlichen Beschreibung von Personen und Körperschaften.

Auch die DDI-Initiative hat eine Reihe von Empfehlungen zu kontrollierten Vokabularen (DDI Controlled Vocabularies, CV) formuliert, um Metadaten wie Erhebungsverfahren,

Datentypen usw. einheitlich zu kodieren. Den Hintergrund der Nutzung und Verarbeitung kontrollierter Vokabulare des DDI Standard durch technische Systeme beschreiben beispielsweise Jääskeläinen, Moschner und Wackerow (2009) in *Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability* (vgl. auch DARIAH-DE).

Zu den standardisierten Vokabularen zählen schließlich fachwissenschaftliche Klassifikationen, die in sozialwissenschaftlichen Erhebungen zur Kodierung entsprechender Variablen genutzt werden. Dazu gehören etwa die *International Standard Classification of Education* (ISCED 2011) der UNESCO, mit der Bildungsabschlüsse klassifiziert werden, oder die *Internationale Standardklassifikation von Berufsgruppen* (ISCO) der International Labor Organisation (ILO) (Jensen 2012). Die Anwendung von kontrollierten Vokabularen in europäischen Archiven und Statistischen Ämtern beschreiben Karjalainen, Kleemola und Jensen (2012).

9.1.1 Der DDI-Metadatenstandard

Die *Data Documentation Initiative* (DDI) entstand 1995 aus einem von dem US-amerikanischen Datenarchiv ICPSR (Inter-University Consortium for Political and Social Research) initiierten Projekt. Ziel war es, die Möglichkeiten einer standardisierten Dokumentation sozialwissenschaftlicher Studien und deren Forschungsdaten zu verbessern und die Erschließung dieser Forschungsressourcen den Möglichkeiten des Webs anzupassen. In die weitere Entwicklung flossen die praktischen Erfahrungen der international vernetzten sozialwissenschaftlichen Datenarchive ein, die zu ersten gemeinsamen Empfehlungen einer standardisierten Studien- und Datendokumentation führten. 2003 wurde die *DDI Alliance* als Mitgliederorganisation gegründet und führte formalisierte Prozesse zur Weiterentwicklung der Initiative ein. In der DDI Alliance sind aktuell mehr als 40 sozialwissenschaftliche Datenarchive, Datenproduzierende, Erhebungsinstitute, Universitäten, kommerzielle Organisationen sowie international tätige Organisationen aus 18 Ländern beteiligt. Langfristiges Ziel der DDI Alliance ist es, den DDI Standard zu einem offiziellen ISO-Norm weiterzuentwickeln.

In Europa findet seit 2009 jährlich die European DDI User Conference (EDDI) statt. Sie ist ein Forum europäischer DDI-Nutzer, die ihre Anwendungen von DDI präsentieren und Fragen und Herausforderungen rund um den Standard diskutieren. Ebenso unterhält die DDI Alliance ein umfangreiches Verzeichnis von Software, mit der DDI-Formate erstellt und verarbeitet werden können. Teil des Publikationsangebotes sind die Best-Practice-Empfehlungen zur DDI-Implementierung. Zur Fortbildung finden regelmäßige Workshops zu DDI statt.

Seit 1995 wurden mehrere Versionen des DDI Standard veröffentlicht. Seit 2012 stehen zwei unterschiedliche Versionen zur Verfügung: DDI-Codebook (DDI-C) und DDI-Lifecycle (DDI-L). Beide Standards sind im *Extensible-Markup-Language*-Format (XML) definiert und über entsprechende Repräsentationen sowohl für Menschen (z.B. bei der Recherche über eine Weboberfläche) als auch für Maschinen (z.B. mit Hilfe von Import- und Exportroutinen) lesbar (vgl. Zenk-Möltgen 2012: 113).

Die zuerst entwickelte Version DDI-C versteht sich als digitales Äquivalent des früher in Papierform produzierten Codebooks. Ein Fokus liegt auf der nachhaltigen Dokumentation von Umfragedaten im Zuge der Archivierung und Nachnutzung von Datensätzen. DDI-Lifecycle setzt einen weiteren Schwerpunkt, indem er die Erfassung und Nachnutzung aller Metadaten ermöglicht, die in den verschiedenen Phasen des Forschungsdatenlebenszyklus oder ggf. auch über verschiedene Studien, Projekte und Organisationen hinweg entstehen. Dazu zählen etwa Metadaten aus den Phasen der Erstellung einer Studienkonzeption, der Datensammlung, der Aufbereitung von Forschungsdaten oder der Datenanalyse sowie solche Metadaten, die auch für die Publikation in Datenkatalogen und die Langzeitarchivierung rele-

vant sind. Die Strukturen beider Versionen des DDI Standard werden im Folgenden kurz vorgestellt.

9.1.2 Der DDI-Codebook Metadatenstandard – DDI-C

Die Version DDI-Codebook dient vorrangig der Dokumentation einfacher Umfragedaten aus Querschnittsstudien (*cross-section study*) und ermöglicht weiterhin die Beschreibung von Mikrodaten, Aggregatdaten und geographischen Angaben. Der DDI-C Standard wurde 2012 als Version DDI 2.5 veröffentlicht. Die wesentlichen Gruppen von Metadaten und ihre wichtigsten Elemente sind im Schaukasten 9.1 auszugsweise vorgestellt.

Schaukasten 9.1: DDI-Codebook – Metadatenstruktur und wesentliche Metadatenelemente (Auszug)

1. Dokumentbeschreibung:
   Beschreibt das DDI-Dokument als Ganzes mit Angaben zu Titel, Autoren, Publikation und Zitation.
2. Studienbeschreibung:
   Die Metadaten dokumentieren Inhalt, Typ und Autoren der Studie sowie den zeitlichen und geographischen Rahmen der Datenerhebung. Es werden Erhebungsmethode, Grundgesamtheit sowie Analyseeinheit und Art der Stichprobe soweit möglich mit kontrollierten Vokabularen beschrieben. Weitere Metadaten erfassen Versionierung, bibliographische Zitation und den dauerhaften Identifikator des Datensatzes. Datenzugang und Nutzungsbedingungen werden ebenfalls dokumentiert.
3. Variablenbeschreibung:
   In diesem Abschnitt werden Variablennamen, Typ und Labels der Variable erfasst. Neben dem Code und den Häufigkeiten der Variable im Datensatz können Hinweise zur Kodierung oder Berechnung von Werten mit Hilfe von Anmerkungen beschrieben werden. Weitere Metadaten beinhalten Fragetexte und Antwortkategorien, Intervieweranweisungen sowie Filterinformationen aus dem Fragebogen.
4. Datensatzbeschreibung:
   Hier werden die Anzahl der Variablen und Erhebungsfälle und Namen, Formate und Versionen der Datendatei(en) dokumentiert.
5. Anderes Material:
   Dieser Abschnitt enthält Informationen und Links zu digitalisierten Kontextdokumenten wie z.B. Fragebögen, Methodenberichte usw. sowie Verweise auf Publikationen, in denen der Datensatz der Studie genutzt wurde.

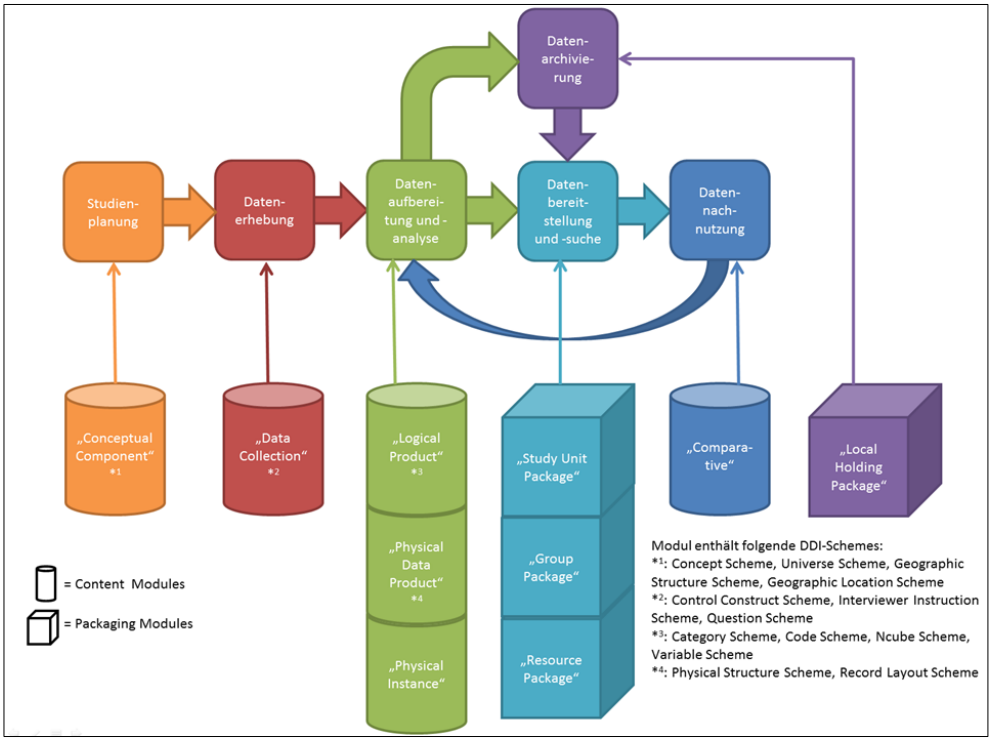
Quelle: Eigene Darstellung in Anlehnung an Jensen (2012: 49)

Diese Metadaten beschreiben die Kontextinformationen zur Studie, die Dokumentation sowie den Datensatz und seine Variablen. Die übergreifenden Informationen zur Studie werden als sogenannte *Studienbeschreibung* über Datenkataloge für differenzierte Recherchen zu Themen einer Studie und zur Bereitstellung von dazugehörigen Daten und Dokumentation angeboten (vgl. Abschnitt 9.2.3). DDI-C erleichtert die Migration der Codebuch-Metadaten in den DDI-Lifecycle Standard. So können etwa DDI-C-Metadaten einer Querschnittstudie in DDI-L überführt werden, sobald zu einer Querschnittsstudie später weitere Erhebungen hinzukommen und die maschinelle Wiederbenutzung vorhandener Metadaten in DDI-L von Interesse ist.

9.1.3 Der DDI-Lifecycle Metadatenstandard – DDI-L

DDI-L basiert auf dem von der DDI Alliance entwickelten Data-Lifecycle (vgl. Kapitel 2.2), der in acht Phasen aufgeteilt ist. Diese müssen jedoch nicht notwendigerweise in chronologischer Reihenfolge durchlaufen werden. Entsprechend wurden in DDI Module entwickelt, die den Data-Lifecycle-Phasen logisch zugeordnet sind und zusammengehörige Metadatenelemente der jeweilig erforderlichen Dokumentation einer Phase zusammenfassen.

Abbildung 9.1: Mögliche Nutzung von DDI-Modulen in den Phasen des Lebenszyklus von Forschungsdaten



Quelle: Eigene Darstellung basierend auf dem Lifecycle-Modell der DDI Alliance (DDI (o.J.): Lifecycle-Modell).

Abbildung 9.1 zeigt die zehn wichtigsten Module und ihre logische Anbindung an die (in der Grafik verkürzt dargestellten) Phasen des DDI-Lifecycle-Modells, die im Folgenden beleuchtet werden.[1] Damit ermöglichen es die Module, einzelne Phasen im Lebenszyklus von Forschungsdaten unabhängig voneinander zu dokumentieren und die entstandenen Metadaten in anderen Phasen wiederzuverwenden. Die Module können auch – unabhängig von der Zuordnung zu einer Phase – danach unterschieden werden,

- ob sie der Erzeugung von Metadaten dienen und diese beinhalten (*Content Modules,* die als runde Form in der Abbildung dargestellt werden) oder
- ob sie den Inhalt der *Content Modules* wiederverwenden und diesen für bestimmte Zwecke strukturieren (*Packaging Modules);* in Abb. 9.1 als eckige Form dargestellt.

Einige der *Content Modules* beinhalten weiterhin eine Reihe von mit Sternchen gekennzeichneten *DDI Schemes*. Das sind Listen von wiederverwendbaren DDI-Metadatenelementen eines bestimmten Typs, wie beispielsweise Fragen (*Question Scheme* im Modul *Data Collection*) oder Variablen (V*ariable Scheme* im Modul *Logical Product*).

Die wesentlichen DDI-L-Module sowie ihre mögliche (wenn auch nicht zwingende) Zuordnung zu den Phasen im Data-Lifecycle werden im Folgenden kurz beschrieben.

1. Phase: Studienplanung und das Modul *Conceptual Components*

1 Alle Namen von Modulen und Teilelementen werden zur besseren Lesbarkeit im Folgenden – im Gegensatz zu DDI Konventionen - nicht zusammengeschrieben.

Dieses Modul dient der Dokumentation von wissenschaftlichen Konzepten (z.B. Arbeitslosigkeit), die von Datenelementen (Variablen) gemessen werden, sowie Grundgesamtheiten (alle Beschäftigten ab 18 Jahre) und geographischen Strukturen (Deutschland) oder Orten, die den dokumentierten Daten zugrunde liegen.

2. Phase: Datenerhebung und das Modul *Data Collection*
   Das Modul dient der Beschreibung der Erhebungsmethode, Erhebungszeiträume und zur Dokumentation besonderer Ereignisse im Rahmen der Datenerhebung. Außerdem werden hier u.a. mehrsprachige Fragetexte und Antwortdomänen (Text, numerisch, Codeschema etc.) sowie Ablauf- und Kontrollstrukturen, Intervieweranweisungen und Codierungsanweisungen des Messinstruments dokumentiert.
3. Phase: Datenaufbereitung und Datenanalyse
   Ggf. ist diese Phase in die Phase Datenarchivierung eingebunden, wenn die Daten in ein Datenarchiv aufgenommen werden. Dabei spielen folgende vier Module eine besondere Rolle (siehe auch Abschnitt 9.3.4):
   - Im Modul *Logical Product* werden die Metadaten zur Struktur der erhobenen Daten abgelegt. Hier sind die Listen der Antwortkategorien und der verwendeten numerischen Codes und die daraus entstehenden Variablen des Datensatzes dokumentiert. Aggregierte Daten von Variablen mit mehreren (N) Dimensionen – sogenannte *Cubes* – oder generell n-dimensionale Datenstrukturen werden mit Hilfe sogenannter *NCubes* erfasst. Variablen und *NCubes* können in Gruppen zusammengefasst und ihre Beziehungen beschrieben werden.
   - Das Modul *Physical Data Product* benennt die physikalischen Eigenschaften der Datenstrukturen, etwa ob die Daten in einem festen, variablen oder Trennzeichen-Format vorliegen. Die Verbindung zu den Variablen aus dem *Logical Product* wird funktional über sogenannte *Data Relationships* gesteuert.
   - Das Modul *Physical Instance* dokumentiert die physikalische Datendatei als Eins-zu-Eins-Relation zu einem konkreten Datenfile (z.B. im SPSS- oder STATA-Format), die die Daten enthält. Das Modul erlaubt auch die Speicherung von Tabellen mit statistischen Auswertungen zu den Variablen.
   - Das spezielle Modul *Local Holding Package* beschreibt im Kontext der Phase Datenarchivierung Metadaten zum Modul *Study Unit Package* (s.u.), die von einem Archiv für die Datenbereitstellung und Datensuche (zusätzlich) erstellt werden.
4. Phasen: Datenbereitstellung und Datensuche (hier gemeinsam ausgewiesen)
   Spätestens zu diesem Zeitpunkt (oft in Verbindung mit einer vorausgehenden Datenarchivierung) kommen die Metadaten folgender Module zum Einsatz:
   - Das Modul *Study Unit Package* beschreibt grundlegende Kontextinformationen der erhobenen Daten. Dazu gehören Metadaten, die der Identifizierung dienen, wie Studiennummer, Persistent Identifier und Zitationsinformationen. Weiterhin werden die räumliche und zeitliche Einordnung der erfassten Daten sowie die abgedeckten Themen dokumentiert. Außerdem erfasst das Modul grundlegende Konzepte der Datenauswahl und -erhebung sowie die abzubildende Grundgesamtheit. Informationen über den Forschungszweck der Studie sowie Angaben zu Forschungsanträgen und deren Finanzierung werden ebenfalls hier dokumentiert.
   - Das Modul *Group Package* erlaubt in diesem Kontext die Vererbung von Basisinformationen, etwa über ein Umfrageprogramm. Werden z.B. im Rahmen dieses Programms wiederholt Daten zu getrennten Zeitpunkten erhoben und dokumentiert, können die grundlegenden Informationen in der Dokumentation der Folgeerhebung wiederverwendet und, wenn notwendig, angepasst werden.
   - Eine wichtige Rolle bei der Wiederbenutzung von Metadaten spielt das *Resource Package*. Dieses Modul ermöglicht die Dokumentation von Elementen, wie z.B. Fragen, Antwortskalen oder Variablendefinitionen, die in standardisierter Form in unterschiedlichsten Studien oder Umfrageprogrammen wiederverwendbar sind.
5. Phase: Nachnutzung
   Das Module *Comparative* ermöglicht im Rahmen der Nachnutzung den paarweisen Vergleich von Elementen wie z.B. Fragen, Variablen, Kategorien und Code-Schemata aus vergleichenden Studien. Gemeinsamkeiten und Unterschiede können anhand der Werte *Identical*, *High*, *Medium*, *Low*, *None* codiert werden.

Die Metadatenelemente aus diesen Modulen können vielfältig über technische kontrollierte Referenzen miteinander vernetzt werden und ermöglichen damit auch eine maximale Wiederverwendung der Dokumentationsteile in den verschiedenen Stadien des Forschungsdaten-Lebenszyklus (Jensen 2012: 48f; Zenk-Möltgen 2012: 114f).

9.2 Metadaten zum Beschreiben und Finden von Studien und Datensätzen

Metadaten auf Studienebene bieten nach Projektende einen Überblick über den Forschungskontext, die Konzeption der Studie und die Methode der Datenerhebung sowie Dokumente zur Datenaufbereitung und Informationen über publizierte Ergebnisse. Sie sind somit auch ein Schlüssel für Forschende, um Daten und notwendige Dokumentationen zu finden und neue wissenschaftliche Fragestellungen mit bereits vorhandenen Daten zu bearbeiten, ohne Daten selbst erheben zu müssen. Diese Nachnutzung von Daten (Data-Sharing) im Rahmen von Sekundäranalysen setzt aber nicht nur voraus, dass die Daten sowohl auffindbar und verständlich dokumentiert sind, sondern erfordert auch deren Zugänglichkeit und technische Nutzbarkeit (s. dazu ausführlich Kapitel 8.1).

Angesichts von national wie international verteilten, heterogenen Datenbeständen, die beispielsweise von Datenarchiven, Repositorien oder Forschungsdatenzentren unterhalten werden, stellt sich auch für Datenproduzierende und Datennutzende in Forschungsprojekten die Frage, wie Studien und Daten ggf. interdisziplinär beschrieben und gefunden werden können. So konstatiert etwa Horstmann (2007: 231):

> Wissenschaftler sind es aber gewohnt, bei der Informationssuche auf nationale oder internationale Datenbestände zuzugreifen und möchten nicht einen lokalen Katalog nach dem nächsten „durchblättern".

Welche Metadatenstandards bei der Entwicklung vernetzter Infrastrukturen zum Nachweis von Forschungsdaten eine wichtige Rolle übernehmen, wird in diesem Beitrag anhand der Standards Dublin Core, DataCite und DDI beschrieben. Im Anschluss wird das Thema *Interoperable Metadaten* an den drei Beispielen behandelt, um zu zeigen, wie ein Set von Kernmetadaten die Erschließung von Forschungsdaten aus unterschiedlichsten Disziplinen ermöglicht.

9.2.1 Metadaten zum Beschreiben von Ressourcen – Beispiel Dublin Core

Dublin Core (DC) ist ein weitverbreiteter Standard zur Beschreibung und Erschließung digitaler Objekte im Internet, der von der *Dublin Core Metadata Initiative* (DCMI) entwickelt und unterhalten wird. Seit dem Gründungsworkshop 1995 hat sich der ursprüngliche Zweck der Metadaten – „improving the discovery of electronic resources on a rapidly growing World-Wide Web*"* – in Richtung der „resource description*"* (DCMI 2011: o.S.) sukzessive erweitert. So sollen DC-Metadaten nicht nur elektronisch verfügbare Objekte, sondern prinzipiell jedes identifizierbare Objekt beschreiben können. Dies können physikalische Dinge, Konzepte, Software aber auch Datensätze und (Daten-)Kollektionen sein, wie sie im (kontrollierten) DCMI Type Vocabulary (2012) beschrieben werden. Vor diesem Hintergrund wird verständlich, dass die DC-Metadaten zur Beschreibung von und Suche nach Objekten nicht nur für Bibliotheken und Museen, sondern im Prinzip für alle wissenschaftlichen Disziplinen von Interesse sind. So können auch Objekte anderer Domänen auf einfache Weise

mit einem Satz von 15 Kernmetadaten beschrieben und online erschlossen werden. Sie werden Dublin Core Metadata Element Set (DCMES 2012) oder kurz DC oder DC Simple genannt und sind als ISO Standard 15836 anerkannt. Schaukasten 9.2 beschreibt (auszugsweise) diese Kernelemente und ordnet sie Metadatentypen (vgl. Abschnitt 9.1) zu, um ihren jeweiligen Zweck zu verdeutlichen.

Schaukasten 9.2: Kernelemente des Dublin-Core-Metadatenschemas (Auszug)

Deskriptive Metadaten

- Title: formeller Name oder Titel des Objektes
- Subject: thematische Einordnung des Objektes, z.B. durch ein Klassifikationssystem
- Description: Abstract zur inhaltlichen Beschreibung des Objektes
- Coverage: räumliche oder zeitliche Zuordnung des Objektes
- Language: Sprache des Inhalts des Objektes

Bibliographische Metadaten

- Creator – Contributor: Produzent des bzw. Mitwirkende an der Erzeugung des Objekts
- Publisher: Instanz, die das Objekt veröffentlicht

Administrative Metadaten

- Rights: rechtliche Eigenschaften des Objektes, z.B. Lizenzen, Zugangsrechte
- Date: relevantes Datum im Lebenszyklus, z.B. wann ein Datensatz geändert wurde

Strukturelle Metadaten

- Identifier: eindeutige Identifizierung des Objekts, z.B. durch ISBN, URL, DOI
- Source: Verweis auf ein Ursprungsobjekt, auf dem das vorliegende Objekt aufbaut
- Relation: Verweis auf ein Objekt, das mit dem beschriebenen Objekt auch in Verbindung steht

Technische Metadaten

- Format: Angabe, wie das Objekt dargestellt ist oder weiterverarbeitet werden kann
- Type: Art und Gattung des Objektes, das durch ein kontrolliertes Vokabular definiert werden kann, wie z.B. Kollektion, Datensatz, Text, Programm, Dienste, Ereignis, physikalisches Objekt, Bildmaterial, Tonmaterial

Quelle: Eigene Darstellung mit Bezug auf DCMES – Dublin Core Metadata Element Set (2012)

Alle Felder sind optional, wiederholbar und können in beliebiger Reihenfolge erscheinen. Sie ermöglichen beispielsweise die Zitation eines Objekts nach unterschiedlichen Zitationsstandards. Um die Zitation und Auffindbarkeit sozialwissenschaftlicher Objekte wie Datensätze und Begleitmaterialien in differenzierter Weise zu ermöglichen, sind semantisch inhaltsreichere, disziplinspezifische Metadaten erforderlich. Solche Metadaten behandelt der nächste Abschnitt.

9.2.2 Metadaten zur Zitation von Forschungsdaten – Beispiel DataCite

DataCite ist ein 2009 gegründetes, stetig wachsendes, internationales Konsortium mit Mitgliedern in Europa, Nordamerika, Asien und Australien. Das Konsortium wurde gegründet, um einheitliche Standards zur Akzeptanz von Forschungsdaten als legitime, eigenständige und zitierfähige wissenschaftliche Leistung weltweit zu etablieren und die Archivierung von sowie den Zugang zu Daten für die Nachnutzung zu fördern.

Das Kernkonzept beruht auf der Nutzung eines *Persistent Identifiers* (PID). Im DataCite-Kontext ist ein PID eine Zuordnung zwischen einer Zeichenfolge und einem Objekt. Objekte können z.B. ein Datensatz, ein Text, eine Audio- bzw. Videodatei, Software, Workflows,

Ereignisse etc. sein. Um PIDs zu erzeugen, nutzt DataCite das System von DOI (Digital Object Identifier), wie im Kapitel 10 näher vorgestellt.

Das aktuelle DataCite-Metadatenschema 4.1 (DataCite 2017: 3f.)

> is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. […] The resource that is being identified can be of any kind, but it is typically a dataset. We use the term 'dataset' in its broadest sense. We mean it to include not only numerical data, but any other research data outputs.

Es umfasst insgesamt neunzehn Hauptelemente (*Properties*), deren Eigenschaften durch weitere Unterelemente (*Subproperties*) spezifiziert werden können.

In Deutschland vergibt die Registrierungsagentur da|ra, die seit 2010 DataCite-Mitglied ist, DOI-Namen für Forschungsdaten und Materialien aus den Sozial- und Wirtschaftswissenschaften. Der da|ra Service hat ein spezielles DDI kompatibles Metdatenschema entwickelt (Koch et al. 2017), um den besonderen Anforderungen an die Beschreibung und differenzierte Erschließung sozial- und wirtschaftswissenschaftlicher Daten gerecht zu werden. Das Schema erweitert das Metadatenschema von DataCite um disziplinspezifische Elemente und besteht aus 32 Hauptelementen (und zusätzlichen Subelementen). Zusammen mit dreizehn kontrollierten Vokabularen können sowohl Forschungsdaten als auch Materialien, die im Forschungsprozess entstanden sind, differenziert beschrieben, zitiert und erschlossen werden (s. Schaukasten 9.3).

Schaukasten 9.3: Notwendige Metadaten zur Registrierung bei da|ra

- Art der Ressource – Beschreibung durch kontrolliertes Vokabular und freien Text
- Titel der Ressource
- Namen der Primärforscher und/oder Namen der Institution, die die Daten erstellt haben
- Publikationsagent – Einrichtung, die die Ressource veröffentlicht
- DOI-Name, der der Ressource zugeordnet ist sowie
- die URL, die auf die Ressource verweist
- Version der publizierten Ressource
- Publikationsdatum der Ressource
- Bedingungen, unter denen die Ressource zugänglich ist

Quelle: Eigene Darstellung mit Bezug auf Koch et al. (2017)

Das erste Element in Schaukasten 9.3 ist von Bedeutung, um die Eigenschaft einer Ressource zu spezifizieren. So kann ein allgemeiner Typ *Datensatz* als spezifischer Typ *Zensusdaten* charakterisiert werden. Darüber hinaus können vielfältige Metadaten genutzt werden, um u.a. einen Datensatz inhaltlich und bibliographisch näher zu beschreiben und Forschende etwa über Thema, Sample und Erhebungsmethoden sowie die zeitliche und geographische Abdeckung der Daten zu informieren. Weiterhin können zum Datensatz gehörende Objekte, z.B. Fragebogen, Methodenbericht oder Syntaxdateien registriert werden. Durch diese disziplinspezifischen Metadaten ist es Forschenden möglich, in Katalogen sehr zielgerichtet (fachspezifische) Daten und Dokumentationen zu interessierenden Fragestellungen zu finden bzw. eigene archivierte Projektdaten zitierfähig zu publizieren.

9.2.3 Metadaten zum Finden sozialwissenschaftlicher Daten – Beispiel DDI

Mit der Entwicklung des DDI Standard begannen die sozialwissenschaftlichen Datenarchive weltweit, die papierbasierten Kataloge ihrer Datenbestände auf elektronische Datenkataloge mit spezialisierten Funktionen zur Suche und Bereitstellung umzustellen.

Zur Beschreibung und Erschließung von wissenschaftlichen Umfragen und zur Archivierung entsprechender Daten wurde das elektronische Format der Studienbeschreibung für

sozialwissenschaftlich orientierte Datenkataloge entwickelt (Bauske 2000). Die Studienbeschreibung dokumentiert anhand der strukturierten, und disziplinspezifischen DDI-Metadatenstandards u.a. Herkunft, Inhalte, Zugänglichkeit und Nutzungsbedingungen der Daten und Dokumentationen. Bibliografische Informationen beschreiben die Studie, die Forschenden sowie deren Publikationen. Methodenbezogene Metadaten umfassen u.a. Auswahlverfahren, Erhebungsmethoden sowie Ort und Zeitraum der Erhebung, wie oben bereits erörtert.

Schaukasten 9.4: DDI-Metadatenelemente der Studienbeschreibung

Bibliographische Angaben

- Studientitel und Studiennummer
- Namen und Institutionen der Primärforscher/innen
- Institution, die die Daten erhoben hat
- Angaben zur Zitation des Datensatzes
- Version des Datensatzes (Nummer, Namen, Datum) und Errata (Korrekturen in den Daten)
- Persistent Identifier (PID) des Datensatzes

Inhalt

- inhaltliche Beschreibung der Studie – Abstract
- Themenklassifikation

Methodologie

- zeitliche und geographische Angaben zur Erhebung
- Grundgesamtheit und Auswahlverfahren
- Typ des Erhebungsverfahrens

Daten und Dokumente

- Format des Datensatzes und Art des Analysesystems
- Anzahl der Einheiten (Fälle) und Variablen
- Datenzugang und Nutzungsbedingungen
- Datensatz (zum Download)
- Codebücher, Fragebogen, Methodenberichte o.ä. zum Download

Veröffentlichungen

- Literatur zur Studie, z.B. Forschungsbericht; Analyseergebnisse

Quelle: Eigene Darstellung in Anlehnung an Jensen (2012: 60f.)

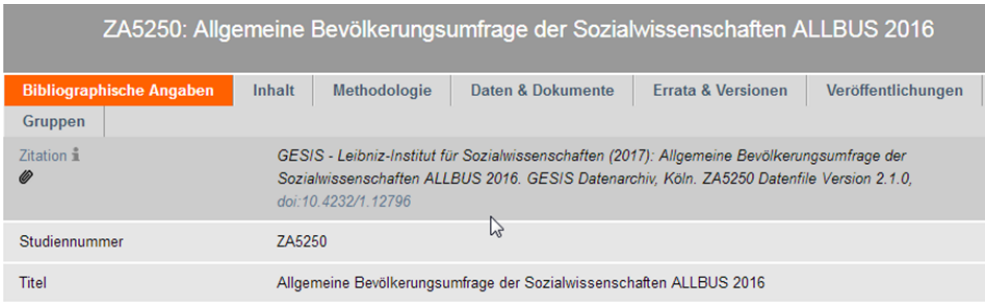
Allgemein betrachtet informieren Metadaten auf Studienebene über den Kontext der Forschungsdaten. In Schaukasten 9.4 sind die wesentlichen Gruppen von Metadatenelementen einer DDI-basierten Studienbeschreibung zusammengefasst. Dabei handelt es sich hauptsächlich um sehr detaillierte deskriptive Metadaten, die durch administrative, strukturelle und technische Metadaten ergänzt werden. Aus Darstellungsgründen wird im Schaukasten auf die gesonderte Kennzeichnung durch diese Typen verzichtet.

Die ausführliche Dokumentation von Studien ist eine Voraussetzung, damit Forschende Daten sowohl finden als auch korrekt interpretieren und nachnutzen können. Für Projekte, die ihre Forschungsdaten und Materialien in einem Datenarchiv sichern und zur Nachnutzung bereitstellen wollen, werden entsprechende Metadaten der Studienebene dort aufbereitet (oftmals auch in englischer Sprache) und in den lokalen Datenkatalog integriert.

Das folgende Beispiel in Abbildung 9.2 stammt aus dem GESIS-Datenbestandkatalog und zeigt auszugsweise Metadaten der archivierten Umfrage ALLBUS 2016 (Studiennummer ZA5250), die zur Nachnutzung bereitsteht. Der Reiter *Bibliographische Angaben* präsentiert Metadaten wie *Zitation und DOI der Daten*, die *Studiennummer* und *Titel* der Studie. Analog zeigen die weiteren Reiter Metadaten zum *Inhalt der Studie*, zur *Methodologie* und zu den downloadbaren *Daten* und *Dokumenten* (Codebuch, Fragebogen usw.). Mit der

Archivierung der Forschungsdaten bei GESIS werden diese bei da|ra registriert und erhalten mit ihrer Veröffentlichung einen DOI-Namen. Informationen zu institutionellen Möglichkeiten der Sicherung, Archivierung und Nachnutzung von Forschungsdaten behandelt Kapitel 7.

Abbildung 9.2: Metadaten der Studie ALLBUS 2016 (ZA5250) im GESIS-Datenbestandskatalog (Ausschnitt)



Quelle: GESIS – Leibniz-Institut für Sozialwissenschaften (2018)

Gleichzeitig werden die Inhalte solcher ‚lokalen' Datenkataloge immer stärker auch grenzüberschreitend mit internationalen Dateninfrastrukturen vernetzt und von unterschiedlichsten Suchmaschinen indexiert. So ist etwa der GESIS-Datenbestandskatalog gemeinsam mit anderen europäischen Archivkatalogen Teil des Datenkatalogs von CESSDA (Consortium of European Social Science Data Archives). Die entsprechenden Metadaten werden darüber hinaus in international vernetzten Online-Ressourcen, wie etwa da|ra oder auf europäischer Ebene im B2Find Katalog (EUDAT) sowie dem Nachweissystem von OpenAire, verwendet. Diese Vernetzung ‚lokaler' Datenkataloge erweitert die Möglichkeiten eines integrierten Nachweises von Forschungsdaten – aber auch von Publikationen und anderen datenbezogenen Forschungsinformationen – aus unterschiedlichen Wissenschaftsdisziplinen.

Dazu ist es notwendig, dass die Metadaten – zumindest für bibliographische Kernelemente wie z.B. den Titel – in den Metadatenschemata der verschiedenen Systeme aufeinander abbildbar sind. D.h., die Metadaten müssen interoperabel sein, was im folgenden Abschnitt thematisiert wird.

9.2.4 Interoperable Metadaten zum Nachweis von Studien und Daten

Die letzten drei Abschnitte haben beispielhaft beschrieben, welche Metadatenstandards für spezifische Zwecke eingesetzt werden können, um Ressourcen auch aus unterschiedlichsten Disziplinen zu finden (z.B. mittels DCMI) und sie zitierfähig im Internet anzubieten (beispielsweise über DataCite, da|ra). Anhand der Metadatenstandards aus dem Bereich der Sozialwissenschaften wurde gezeigt, wie Metadaten es erlauben, Studien und Datensätze dieser Domäne sowohl in sozialwissenschaftlichen als auch in transdisziplinären Datennachweissystemen zu finden (da|ra, DDI, B2Find, OpenAire). Die ausgewählten Standards zeigen aber auch, dass sie Forschungsdaten mit ihren Metadatenelementen unterschiedlich tief beschreiben und erschließen können. Das bedeutet einerseits, dass disziplinübergreifende Standards erforderlich sind, um Datenressourcen unterschiedlichster Herkunft z.B. in interdisziplinären Informationsplattformen zu finden. Anderseits wird die Notwendigkeit kleinteiliger, disziplinspezifische Metadaten deutlich, um mit ihrer Hilfe z.B. die relevanten sozialwissenschaftlichen Daten für eine spezifische, sekundäranalytische Fragestellung aus der Masse von Datenangeboten herauszufiltern (Jensen/Katsanidou/Zenk-Möltgen 2011).

Die Verwendung solcher Metadaten erfordert es auch, dass sie sowohl von den Nutzenden eines Datenkatalogs als auch von diversen technischen Systemen verstanden, genutzt und verarbeitet werden können. D.h., Metadatenstandards sollten so entwickelt werden, dass sie von Recherchesystemen verarbeitet und erschlossen und auch zwischen solchen Systemen ausgetauscht werden können. Dabei ist sicherzustellen, dass eine Ressource, wie z.B. ein sozialwissenschaftlicher Datensatz, in unterschiedlichen Systemen mit möglicherweise unterschiedlichen Metadatenstandards bzw. Elementen beschrieben und verstanden werden kann.

Damit verschiedene Metadatenstandards untereinander kompatibel sind, müssen grundlegende Metadatenelemente, wie z.B. Name und Quelle eines Objektes interoperabel sein, d.h. zwischen verschiedenen Systemen – möglichst ohne Informationsverlust – ausgetauscht werden können. Im Kontext von Metadaten wird dabei unterschieden zwischen (Rühle o.J.: 5):

- struktureller Interoperabilität – die Standards beruhen auf einem gemeinsamen Datenmodell,
- syntaktischer Interoperabilität – die Metadaten werden in einem Format, wie z.B. XML, kodiert,
- semantischer Interoperabilität – die verwendeten Metadatenelemente haben die gleiche Bedeutung.

Angesichts der Vielfalt von Metadatenstandards, die die besonderen inhaltlichen Anforderungen an Metadaten zur Dokumentationen von Daten in den unterschiedlichsten Disziplinen wiederspiegeln, ist es nicht möglich, diese Arten von Interoperabilität vollständig zu realisieren. Vielmehr bestimmt der Zweck eines Datenkataloges, welche und wie viele Metadaten interoperabel sein müssen. So ist es zum Nachweis von lokalen Datenbeständen in übergreifenden Katalogen ausreichend, wenn eine geringe Zahl interoperabler Kernmetadaten (vgl. Schaukasten 9.2 zum Dublin Core Standard) genutzt wird.

Ein gängiges Mittel, um Metadaten eines Ausgangs- bzw. Quellsystems in einem neuen Zielsystem abzubilden, ist das sogenannte *Mapping*. Dazu werden relevante Metadatenelemente des Quellsystems auf entsprechende Elemente des Zielsystems abgebildet. Um dabei sinnvolle Ergebnisse zu erzielen, ist es erstens notwendig, dass das Mapping semantisch korrekt ist. Dazu werden sogenannte Mappingtabellen bei der Spezifikation eines Metadatenschemas definiert. Zweitens muss die technische Abbildung eines Elements aus dem Quellsystem auf ein anderes Element des Zielsystems syntaktisch möglich sein (z.B. durch Nutzung des XML-Formats) und regelkonform erfolgen.

Üblicherweise werden Mappings mit Elementen anderer Metadatenstandards in der jeweiligen Spezifikation des Standards des Quellsystems systematisch dokumentiert, wie z.B. im da|ra-Metadatenschema Version 4 (Koch et al. 2017: 69ff.). Das Mapping von Metadaten eines Metadatenschemas auf verschiedene Softwaresysteme beschreiben Akdeniz und Zenk-Möltgen (2017: 55ff.). Wie das Mapping bzw. der Austausch von Metadaten aus den Standards von Dublin Core, DataCite und DDI für unterschiedliche Anwendungszecke erfolgt, sei zum Abschluss kurz erwähnt. Dabei spielen die 15 Kernelemente des Dublin Core Standard eine zentrale Rolle (vgl. Schaukasten 9.2).

Sollen beispielsweise DDI-Metadaten einer Studienbeschreibung eines lokalen Datenkatalogs als Quellsystem in ein transdisziplinäres Datenportal wie der B2Find-Katalog von EUDAT als Zielsystem aufgenommen werden, werden die Metadaten automatisiert abgefragt, sodass Forschungsdaten aus verteilten Quellsystemen im Portal gefunden werden können. Bei der Abfrage von Metadaten eines Datenservices, auch Harvesting genannt, wird bevorzugt die Schnittstelle OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) eingesetzt. Dieses Protokoll setzt die 15 Kernelemente von Dublin Core als verpflichtende Metadatenspezifikation (vgl. Spezifikation OAI-PMH Dublin Core; kurz oai_dc) ein. Diese Kernelemente stellen den kleinsten gemeinsamen Nenner an Informationen über ein Objekt dar, die zwischen unterschiedlichen Systemen mittels OAI-PMH ausgetauscht werden

müssen. Das Protokoll erlaubt jedoch auch die Verwendung anderer Metadatenstandards. So verwendet etwa da|ra für seine Metadaten des OAI-PMH Providers zusätzlich zwei da|ra-spezifische Formate, basierend auf dem DDI-Lifecycle-Format und dem DataCite-Format (da|ra OAI-PMH).

9.3 Metadaten zur Dokumentation von Variablen und Fragen

Datenstrukturen in den quantitativen Sozialwissenschaften bestehen in der Regel aus rechteckigen Tabellen, in denen eine Liste von Subjekten in den Zeilen und die Merkmale in den Spalten, die dann Variablen genannt werden, eingetragen sind. So sind z.B. Umfragedaten oft mit einer Zeile pro befragte Person angegeben und in den Spalten die Antworten auf die Fragen des Fragebogens. Ein anderes Beispiel sind Aggregatdaten, etwa wenn Länder in den Zeilen angegeben werden, und Werte für das Bruttosozialprodukt oder die Arbeitslosenquote in der Spalte.

Die Nutzung derartiger Datenstrukturen, etwa in Sekundäranalysen, setzen die genaue Kenntnis einzelner Merkmale voraus. Dazu gehören u.a. die Bedeutung einzelner Variablen, ihre Maßeinheiten und Informationen über das Zustandekommen der Messungen. Hierfür werden in der Regel die Variablen mit kurzen Überschriften (*Labels*) sowie mit kurzen Erklärungen, die z.B. einen Fragetext enthalten können, versehen.

Weitere Informationen können sich z.B. auf den Typ der Daten (etwa Textdaten, numerische Daten, Datumsformate), das Messniveau, die Einheiten der Messung und die gültigen und ungültigen bzw. fehlenden Antwortwerte beziehen (*Missing Values*).

Je nach verwendetem Datenformat lassen sich unterschiedliche Metadaten zusammen in einer Datei mit den eigentlichen Umfragedaten ablegen. Eine Übersicht zu den gängigsten Formaten und ihrer Möglichkeiten für Metadaten im Bereich sozialwissenschaftlicher Datendokumentation und -analysen bietet der folgende Abschnitt. Wir beginnen damit, eher einfachere Strukturen mit wenigen Metadaten zu erörtern und beziehen dann komplexere Strukturen mit ein, die reichhaltigere Möglichkeiten der Dokumentation bieten.

9.3.1 Dokumentation einfacher Datenstrukturen – Beispiel CSV

Einfache Datenstrukturen werden oft im CSV-Format in einer Textdatei abgelegt. CSV steht für *comma-separated values* und wird von einer Vielzahl an Programmen für den Import oder Export einfach strukturierter Informationen verwendet (vgl. Spezifikation RFC4180). CSV-basierte Daten sind immer in einer rechteckigen Datenstruktur angeordnet, da die Daten in Zeilen (befragte Person) und Spalten (Variablen) organisiert werden. Die Zeichenkodierung der Werte ist nicht festgelegt, in der Regel wird aber 7-bit ASCII oder ein Unicode-Zeichenformat wie UTF-8 (vgl. Spezifikation RFC3629) verwendet.

Zur Strukturierung und Abgrenzung der Daten müssen in der Textdatei spezielle Steuerzeichen eingesetzt werden. Deren Kenntnis ist notwendig, um die Dateistruktur formal zu prüfen, z.B. ob die Anzahl von Spalten und Zeilen mit der Anzahl von befragte Person und Variablen übereinstimmt. Auch sollten die erfassten Einträge (Werte) des Datensatzes auf logische Konsistenz mit den Vorgaben des Codeplans kontrolliert werden, bevor die Daten für Analysen genutzt werden.

Die Trennung von Datensätzen (Zeilen) wird in der Regel durch einen Zeilenumbruch gekennzeichnet, dabei kann es Unterschiede zwischen Betriebssystemen geben: In Windows

wird z.B. ein Zeilenumbruch durch die Steuerzeichen *carriage return* (Rückkehr an den Zeilenanfang) und *line feed* (Zeilenvorschub) kodiert, unter Unix ist nur ein *line feed* üblich (vgl. Trost Media 2018).

Die Trennung von Datenfeldern, also den Spaltenwerten, wird durch ein Komma codiert, kann aber auch durch Steuerzeichen wie Semikolon, Tabulatorzeichen (in dem Fall wird das Format auch als TSV bezeichnet) oder andere erfolgen. Die zur Trennung verwendeten Steuerzeichen dürfen nicht als Zeichen in den Werten der Felder selbst vorkommen, z.B. in Form von Variablenwerten mit Dezimalstellen, da dies sonst zur fehlerhaften Trennung führen würde. Aus diesem Grund können auch Feldbegrenzungszeichen (z.B. doppelte oder einfache Anführungszeichen) verwendet werden. Innerhalb dieser Feldbegrenzungszeichen können dann die ansonsten zur Trennung reservierten Zeichen auch in Feldern genutzt werden.

Das CSV-Format kennt keine Metadaten außer optional eine erste Zeile, die die Spaltennamen enthalten kann. Die in den Spalten abgelegten Zahlen-, Datums- und Uhrzeitwerte sind oft länderspezifisch und nicht im Vorhinein festgelegt. Daher sind häufig Zusatzinformationen nötig, um die Inhalte einer CSV-Datei korrekt interpretieren zu können. Sind diese Metadaten nicht bekannt, kann mit Hilfe der vorliegenden Datei nur noch versuchsweise auf die verschiedenen Möglichkeiten von Werten geschlossen werden. Dieser Sachverhalt unterstreicht die Bedeutung einer transparenten und vollständigen Datendokumentation.

Eine Bearbeitung des CSV-Formats ist mit einfachen Texteditoren, wie dem Notepad oder Notepad++, möglich und erlaubt es, die Datensätze auf korrekte Erfassung der Werte in den Spalten und die Metadaten in der Kopfzeile (Spaltennamen) zu kontrollieren. In den Editoren werden die Zeichenkodierung ebenso wie vorhandene Trennzeichen und Textbegrenzungszeichen angezeigt. Tabellenkalkulationen wie Microsoft Excel und Apache Openoffice Calc können CSV lesen und schreiben. Statistikprogramme wie R, SPSS oder STATA können ebenfalls CSV lesen und schreiben, dazu wird dann mit Hilfe eines Import- bzw. Export-Dialogs die Formatierung abgefragt. Da viele Programme zur Erstellung von Umfragen den Export der Daten im CSV-Format erlauben (zum Vergleich von Umfragesoftware siehe z.B. Siegl (o.J.)), wird es häufig als Ausgangsformat verwendet, um weitere Formate zu erzeugen, z.B. RDF zur Erzeugung von Linked Open Data (Lebo/Williams 2010). Im Falle von hierarchischen Datenstrukturen werden eher andere Datenformate verwendet, die ebenfalls textbasiert sind, etwa JSON (JavaScript Object Notation) oder XML (Extensible Markup Language), auf die hier nicht näher eingegangen werden soll (vgl. hierzu Nurseitov et al. 2009).

9.3.2 Metadaten zur Variablendokumentation – Beispiel Statistikprogramme

Statistikprogramme dienen in den Sozialwissenschaften der numerischen Auswertung von erhobenen Forschungsdaten. Sie umfassen zahlreiche Werkzeuge zum Forschungsdatenmanagement, zur Datenmodellierung sowie der Ergebnispräsentation. Die Grundfunktionen werden u.a. durch Module zur Automatisierung (Programmcode, Syntaxeditoren) oder spezialisierte Analysemethoden ergänzt. Im Folgenden stehen die Datenmanagementfunktionen im Vordergrund, die der Erfassung von Metadaten zu Datensätzen und Variablen dienen. Diese werden im Laufe unterschiedlicher Phasen des Forschungsdatenmanagements durch eine Reihe von Einzelschritten bearbeitet.

Daten einer empirischen Erhebung müssen zunächst definiert werden, bevor sie analysiert werden können. Dazu wird auf Basis des Messinstruments, z.B. eines Fragebogens, eine Datendefinition erzeugt. Diese umfasst die Erstellung von Variablendefinitionen und Kodierungsschemata eines Datensatzes. Dabei wird ein erhobenes Merkmal als Variable im Analyseprogramm abgebildet und durch Variablenattribute beschrieben. Im Verlauf der weiteren

Datenaufbereitung bis zur Erstellung eines vollständigen Analysedatensatzes können schrittweise zusätzliche Arten von Variablen z.B. für befragte Personen, Länder oder Zeitpunkte (administrative Variablen) oder zur Gruppierung von Einkommen oder anderen Indices (inhaltliche Variablen) definiert werden. Im Verlauf von Datenanalysen können Variablen ergänzt, modifiziert oder harmonisiert werden.

Wesentliche Metadaten zu Variablen können von allen gängigen komplexeren Statistikprogrammen wie etwa SAS, SPSS, STATA oder R erzeugt werden (zum Vergleich von Statistiksoftware vgl. inwt-statistics.de). Dabei handelt es sich vor allem um deskriptive und technische Metadaten, wie im Schaukasten 9.5 zusammengefasst.

Metadaten in Statistikprogrammen liefern in der Regel wichtige Informationen, die im Laufe der Analyse der Daten herangezogen werden. Um hier eine Wiederholung und Überprüfung von vorgenommenen Veränderungen nachvollziehen zu können, ist die Verwendung von Kommandoskripten für die Festlegung bzw. zur Veränderung der Datendefinitionen empfehlenswert. Dies wird z.B. in SPSS durch Syntax-Dateien und in STATA durch die sogenannten do-Files geleistet. Derartige Skriptdateien erlauben es, von einem Ausgangsdatensatz immer wieder den Ablauf einer Datendefinition, -bereinigung oder -korrektur nachzuvollziehen und so überprüfbar zu machen (vgl. Ebel 2015). Dies gilt insbesondere, wenn die Daten aus einfacheren Datenstrukturen, etwa CSV, importiert werden.

Schaukasten 9.5: Deskriptive und technische Metadaten in gängigen Statistikprogrammen

Deskriptive Metadaten:

- der Datensatzname und evtl. ein Label,
- die Variablennamen und ein Variablenlabel für das erhobene Merkmal,
- die möglichen Werte und ihre Value Labels zur Beschreibung der Merkmalsausprägungen
- und eine Definition fehlender Werte.

Technische Metadaten:

- den Datentyp (numerisch, alphanumerisch, etc.),
- den Bereich gültiger Werte oder die Darstellung von Dezimalstellen.

Quelle: Eigene Darstellung

Für eine vollständige Erfassung aller Metadaten auf Variablenebene sind die Statistikprogramme jedoch in der Regel nicht geeignet. So kann beispielsweise für eine Variable zur Arbeitslosenquote nicht die methodische Grundlage der Daten und das Prozedere zur Ermittlung einer Arbeitslosenquote anhand standardisierter Metadaten erfasst werden. Im Label könnte lediglich etwa eine Definition oder ein Verweis auf ein zugrunde liegendes Konzept genannt werden. So macht es einen wesentlichen Unterschied, ob eine Befragung oder eine amtliche Meldung zur Berechnung der Quote verwendet wurde. Eine anschauliche Beschreibung der methodischen Grundlagen in diesem Kontext geben z.B. Bersheim, Oschmiansky und Sell (2014). Die Erfassung solcher methodischen Grundlagen durch Metadaten behandelt Abschnitt 9.2.3. Darüber hinaus zählt es zur guten wissenschaftlichen Praxis, diese Kontextinformationen für die Nachnutzung der Daten systematisch und nachvollziehbar in einem Methodenbericht darzustellen (Watteler 2010).

Bei Variablen, die auf Umfragen beruhen, wird manchmal versucht, den Fragetext im Variablenlabel zu dokumentieren. Dies ist jedoch durch die Begrenzung der Zeichenlänge und fehlende Formatierung nur z.T. möglich. Darüber hinaus gehen längere Texte in den Labels bei vielen Analysetabellen in der Darstellung wieder verloren. Daher ist es sinnvoller, dort kurze Bezeichnungen zu verwenden (vgl. entsprechende Leitlinien und Regeln in Netscher/Eder 2018; Ebel/Trixa 2015; Jensen 2012).

Besonders im Fall kodierter Variablen aus Fragen in Interviews ist eine Dokumentation der Codes und ihrer Zugehörigkeit zu den Antwortvorgaben zentral für eine richtige Interpretation der Analysen. Da während der Datenbereinigung und einer eventuellen Integration in einen Datensatz mit mehreren Wellen oder mehreren Populationen auch oft die Kodierungen verändert werden, muss auf eine Harmonisierung der veränderten Antwortvorgaben aus dem Fragebogen mit den Value Labeln im Datensatz geachtet werden. Ein Beispiel dazu wäre eine Integration zweier Datensätze, von denen einer die Verwendung von *1* als Code für *Ja* und *0* als Code für *Nein* hat und der andere eine Verwendung von *2* als Code für *Ja* und *1* als Code für *Nein*. Entsprechend müssen die Codes angepasst und ggf. die Value Label des integrierten Datensatzes korrigiert werden. Auch für solche Fälle empfiehlt sich die Verwendung von Skriptdateien zur Dokumentation der durchgeführten Veränderungen. In diesen können die Berechnungsvorschriften als Kommentare abgelegt werden und erlauben so immerhin die Nachvollziehbarkeit durch andere Forschende. Ein solches nicht standardisiertes Vorgehen ist jedoch kein Weg, um eine computergestützte Nachnutzung zu ermöglichen.

9.3.3 Austausch von Metadaten zu Umfragedaten – Beispiel Triple-S

Bereits in den 1990er Jahren wurde das Format Triple-S für den Austausch von Metadaten zu Umfragedaten entwickelt (Hughes/Jenkins/Wright 2000). Triple-S bezeichnet das XML-Format, das einen einfachen Austausch zwischen vielen Programmen in der Umfrageforschung ermöglicht. Die Spezifikation von Triple-S folgt einer Document Type Definition (DTD). Im Jahr 2006 wurde die Version 2.0 des Formats publiziert und seit 2017 ist die Version 3.0 verfügbar, die u.a. eine UTF-8-Kodierung der Daten und eine HTML-Formatierung der Metadaten erlaubt (Triple-S 2017). Mit Hilfe der Angaben im XML-File können z.B. auch Daten im CSV-Format dokumentiert werden. Die möglichen Metadaten von Triple-S sind Variablennamen und -labels, Positionen und Identifier, Antwortwerte und -codes. Ein Beispiel für die Metadaten einer Variablen in Triple-S zeigt Abbildung 9.3.

Zahlreiche Online-Umfrageprogramme (vgl. websm.org 2018) erlauben es, die Metadaten im Triple-S-Format auszugeben, so z.B. LimeSurvey, die Produkte von NIPO, SnapSurveys, SensusWeb, IdSurvey und viele andere. Für Triple-S gibt es einen Validierungsservice und einige freie Tools zur Anwendung mit den Programmen Quantum und IBM SPSS Statistics (vgl. u.a. Export SPSS to Triple-S). Auch für die Statistiksoftware R existiert ein Paket *sss*, mit dem Triple-S verarbeitet werden kann. Die allermeisten mit Triple-S kompatiblen Programme verwenden die Version 1.1 oder 2.0 der Spezifikation. Der Vorteil bei der Verwendung der weiteren Metadaten in Analyseprogrammen besteht u.a. darin, dass für Tabellen ausführlichere Variableninformationen verwendet werden können, z.B. Fragetexte und Antworttexte. Bei der Übertragung werden zu lange Angaben jedoch oft abgeschnitten. Insgesamt werden Fehler bei der Interpretation, etwa wenn die Bedeutungen der Werte verwechselt werden, jedoch durch den Austausch in einem standardisierten Format unwahrscheinlicher, wie bereits in Abschnitt 9.3.2 erörtert.

Abbildung 9.3: Beispiel für die Metadaten einer Variable in Triple-S

```
<variable ident="24" type="single">
    <name>tv1</name>
    <label>Wie viel Zeit verbringen Sie an einem gewöhnlichen Werktag
        insgesamt damit fernzusehen?</label>
    <position start="453" finish="453"/>
    <values>
        <range from="1" to="8"/>
        <value code="1">gar keine Zeit</value>
        <value code="2">weniger als 1/2 Stunde</value>
        <value code="3">1/2 bis zu 1 Stunde</value>
        <value code="4">mehr als 1, bis zu 1 1/2 Stunden</value>
        <value code="5">mehr als 1 1/2, bis zu 2 Stunden</value>
        <value code="6">mehr als 2, bis zu 2 1/2 Stunden</value>
        <value code="7">mehr als 2 1/2, bis zu 3 Stunden</value>
        <value code="8">mehr als 3 Stunden</value>
    </values>
</variable>
```

Quelle: Eigene Darstellung

9.3.4 DDI-Metadaten zur Dokumentation von Variablen und Fragen

Seit einigen Jahren ist das DDI-Format der De-facto-Standard für die Dokumentation von Umfragedaten aus den Sozialwissenschaften. Es stellt im Vergleich zu den bisher geschilderten Standards die reichhaltigsten Möglichkeiten der Metadatendokumentation zur Verfügung. Einen generellen Überblick zu DDI beschreibt Abschnitt 9.1. Die Verwendung von Metadaten zum Finden sozialwissenschaftlicher Daten thematisiert Abschnitt 9.2.3. Im Folgenden werden besonders für die Ebenen der Variablen und Fragen die umfangreichen Möglichkeiten von DDI dargestellt.

Im DDI-Codebook-Format orientiert sich die Dokumentation an der Struktur eines Datensatzes (s. Abbildung 9.4). In der Baumstruktur der XML-Elemente findet sich die Dokumentation von Variablen im zweiten Unterelement des Wurzelements (*codeBook*) im Element für die Datensatzbeschreibung (*dataDescr*). Unterhalb dieses Elements gibt es eine Liste der Variablen (jeweils das Element *var*). Zu jeder Variable gibt es eine Kurzbeschreibung (im Element *labl*) und ein Element für die Fragedokumentation (*qstn*) durch weitere Unterelemente. Im Beispiel sind der wörtliche Fragetext (*qstnLit*) und eine Sprunganweisung (*forward*, möglich wäre auch *backward*) enthalten; die Sprunganweisung referenziert eine nachfolgende Frage im Zusammenhang mit einer Filterfrage. Die Nummer der Frage im Fragebogen wird in einem Attribut zum Frageelement (*seqNo*) angegeben. Weitere Möglichkeiten für Unterelemente zur Fragedokumentation sind z.B. Interviewer-Anweisungen (*ivuInstr*) oder Vor- und Nachfragetext (*preQTxt*, *postQTxt*). Die Antwortvorgaben zur Frage finden sich in weiteren Unterelementen zur Variable (*catgry*), nicht jedoch – wie zu erwarten gewesen wäre – unterhalb der Frage. Jede Antwortvorgabe hat Angaben zum Wert (*catValu*), Label (*labl*), Antworttext (*txt*) und jeweils zu fehlenden Werten (Attribut *missing* mit Wert Y=yes oder N=no). Weitere Angaben zur Variable beziehen sich z.B. auf Informationen zur Ableitung aus einer anderen Variable (*derivation*), auf das technische Format (*varFormat*) oder Anmerkungen (*notes*).

Abbildung 9.4: Beispiel einer Variable und Frage im DDI-Codebook-Format (leicht gekürzt)

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE codeBook SYSTEM "http://www.icpsr.umich.edu/DDI/Version2-0.dtd">
<codeBook ID="ZA3811">
    <stdyDscr>
        <citation><titlStmt>
        <titl>EVS - European Values Study 1999 - Integrated Dataset</titl>
        <IDNo>ZA3811</IDNo></titlStmt></citation>
        <stdyInfo><sumDscr><nation ID="S39" abbr="all countries"></nation></sumDscr></stdyInfo>
    </stdyDscr>
    <dataDscr>
        <var ID="VAR401" name="v293">
            <labl>having steady relationship (Q86)</labl>
            <qstn ID="SQ584" seqNo="Q.86" sdatrefs="S39">
                <qstnLit>Whether you are married or not: Do you live in a stable relationship
                        with a partner?</qstnLit>
                <forward>if "no" - go to 88</forward>
            </qstn>
            <catgry ID="AV40881" missing="Y">
                <catValu>-1</catValu>
                <labl>don't know</labl>
                <txt ID="SA42073" sdatrefs="S39">don't know</txt>
            </catgry>
            <catgry ID="AV993" missing="N">
                <catValu>1</catValu>
                <labl>yes</labl>
                <txt ID="SA1206" sdatrefs="S39">yes </txt>
            </catgry>
            <catgry ID="AV994" missing="N">
                <catValu>2</catValu>
                <labl>no</labl>
                <txt ID="SA1207" sdatrefs="S39">no </txt>
            </catgry>
            <derivation>
                <drvdesc>Iceland:
In the Icelandic questionnaire v293, v294 and v296 were grouped into one question as follows:
'Are you now:  1 - Married, 2 - Living with a partner (but not married), 3 - Divorced, 4 - Separated,
5 - Widowed, 6 - Single (never been married)'. V293 and v294 were reconstructed from original v296, v297.
Turkey:
In the Turkish questionnaire v293, v294 and v295 were not included to avoid possible offences.
V296 was modified to give most of the information asked for in the Master Questionnaire.
V293 and v294 were reconstructed from original v296, v297. V295 could not be reconstructed.</drvdesc>
            </derivation>
            <varFormat type="numeric" formatname="F2.0" schema="SPSS"/>
            <notes type="NoteNote">Trend question: EVS 2008 and EVS 1999.
(Modified trend: interviewer instruction, question wording).</notes>
        </var>
    </dataDscr>
</codeBook>
```

Quelle: Eigene Darstellung

Eine Wiederbenutzung von Fragen durch mehrere Variablen wäre durch die Verwendung des Attributs *qstn* (nicht im Beispiel gezeigt) im Element *qstn* möglich. Dieses Attribut enthält eine Referenz auf die ID einer Frage, sodass die untergeordneten Textinhalte nicht wiederholt aufgeführt werden müssen. Eine von verschiedenen Variablen verwendete Frage muss so nur einmal komplett dokumentiert werden und kann von allen Variablen referenziert werden.

Im DDI-Lifecycle-Format stehen verschiedene Module für die Dokumentation bereit (vgl. Abbildung 9.1). Die Dokumentation von Fragen ist im Modul *DataCollection* vorgesehen. Für die Dokumentation von Datensätzen und ihren Variablen sind es die Module *LogicalProduct* für logische Informationen, *PhysicalDataProduct* für konkretere Informationen zum Datensatz und *PhysicalInstance* für Dateiinformationen. Die Ablage der Elemente dieser Module in einem *ResourcePackage* erlaubt die Wiederverwendung verschiedener Elemente wie Fragen, Antwortskalen, Codelisten und Textstatements in unterschiedlichen Studien, Fragebögen oder Datensätzen. Das Beispiel in Abbildung 9.5 listet Elemente einer Frage auf, wie sie im *ResourcePackage* abgelegt werden: Eine Abfolge von Texten im Fragebogen (*Sequence*) enthält eine Referenz auf einen Fragebogentext (*StatementItem*) und eine konkrete Frage (*QuestionConstruct*). Dabei beinhaltet der Fragebogentext im Beispiel

eine Sprunganweisung zu einer anderen Frage. Die konkrete Frage enthält einen Verweis auf das Frage-Item (*QuestionItem*). Hier kann eine Referenz auf eine Anweisung (*InterviewerInstruction*) eingefügt werden (nicht im Beispiel). Das eigentliche Frage-Item wiederum dokumentiert die Fragenummer (in *QuestionItemName* mit dem Attribut *context* gleich *questionNumber*) und den wörtlichen Fragetext (in *LiteralText*), sowie mit Hilfe einer Referenz die möglichen Antwortvorgaben (*CodeListReference*). In der referenzierten *CodeList* sind die Antworten und ihre Werte dokumentiert (nicht in Abbildung 9.5 enthalten).

Die DDI-Lifecycle-Dokumentation einer Studie (*StudyUnit*) nutzt diese Struktur, indem ein Verweis von der Datenerhebung (*DataCollection*) auf das Erhebungsinstrument (*Instrument*) zeigt, welches wiederum die im Beispiel gezeigte Abfolge von Texten im Fragebogen (*Sequence*) referenziert. Ähnlich wird bei der Dokumentation der Variablen vorgegangen. Unter den logischen Informationen zum Datensatz (*LogicalProduct*) einer Studie wird eine Referenz auf die Liste der Variablen (*VariableScheme*) gesetzt. Die Verknüpfung zwischen Variablen und Fragen erfolgt unterhalb der Variable mit Hilfe eines Verweises auf ein Frage-Item.

Abbildung 9.5: Beispiel einer Frage im DDI-Lifecycle-Format (gekürzt)

```
<dc:Sequence>
 <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Sequence_Q375</r:ID><r:Version>1.0.0</r:Version>
  <dc:ControlConstructReference>
   <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Statement_Q375</r:ID><r:Version>1.0.0</r:Version>
   <r:TypeOfObject>StatementItem</r:TypeOfObject>
  </dc:ControlConstructReference>
  <dc:ControlConstructReference>
   <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Construct_Q375</r:ID><r:Version>1.0.0</r:Version>
   <r:TypeOfObject>QuestionConstruct</r:TypeOfObject>
  </dc:ControlConstructReference>
 </dc:Sequence>

  <dc:StatementItem>
   <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Statement_Q375</r:ID><r:Version>1.0.0</r:Version>
   <dc:DisplayText>
    <dc:LiteralText><dc:Text xml:lang="en">if "no" - go to 88</dc:Text>
    <!-- CMM Portfolio V1: 6.3.2.3 Statementitem -->
    </dc:LiteralText>
   </dc:DisplayText>
  </dc:StatementItem>

  <dc:QuestionConstruct>
   <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Construct_Q375</r:ID><r:Version>1.0.0</r:Version>
    <r:QuestionReference> <!-- CMM Portfolio V1: 4.9.12  Association Question -->
     <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Q375</r:ID><r:Version>1.0.0</r:Version>
     <r:TypeOfObject>QuestionItem</r:TypeOfObject>
    </r:QuestionReference>
  </dc:QuestionConstruct>

  <dc:QuestionItem>
   <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_Q375</r:ID><r:Version>1.0.0</r:Version>
   <r:UserID typeOfUserID="QuestionID">Q375</r:UserID><!-- CMM Portfolio V1: 6.1 Question-ID -->
   <dc:QuestionItemName context="questionNumber"><r:String>Q.86</r:String></dc:QuestionItemName>
   <!-- 6.2.1.1 Question Number -->
   <dc:QuestionItemName context="title-english"><r:String xml:lang="en">F375: ZA3811-V293-do you live
      in stable relationship with partner (Q86)</r:String></dc:QuestionItemName>
      <!-- CMM Portfolio V1: 6.3.1.1 Question Label -->
   <dc:QuestionText>
    <dc:LiteralText><dc:Text xml:lang="en">Whether you are married or not: Do you live in a stable
      relationship with a partner?</dc:Text> <!--CMM Portfolio V1: 6.3.2.2  Question Text -->
    </dc:LiteralText>
   </dc:QuestionText>

   <dc:CodeDomain>
    <r:CodeListReference> <!-- CMM Portfolio V1: 6.2.2   Reference to Codelist -->
     <r:Agency>de.gesis</r:Agency><r:ID>ZA3811_v293_CodLis</r:ID><r:Version>1.0.0</r:Version>
     <r:TypeOfObject>CodeList</r:TypeOfObject>
    </r:CodeListReference>
   </dc:CodeDomain>
  </dc:QuestionItem>
```

Quelle: Eigene Darstellung

Zusätzlich kann eine Dokumentation einer abgeschlossenen Studie die Grundlage für die Erstellung eines neuen Fragebogens sein. Die dann erstellten Verweise dokumentieren auch systematisch, aus welchen vorherigen Studien oder Fragebögen die Fragen übernommen wurden, sodass eine größere Transparenz der Datenkollektion gegeben ist. Eine dabei nutzbare Funktionalität bildet die Versionierung von Elementen in DDI-Lifecycle: Soll beispielsweise der Text einer Frage in einer neuen Verwendung leicht angepasst werden, wird eine neue Version der Frage dokumentiert, die nur den geänderten Text enthält, jedoch auf die anderen Elemente in ihrer ursprünglichen Version verweist. Hierdurch ergibt sich eine systematische Dokumentation von Veränderungen und Anpassungen des Erhebungsinstruments (vgl. Recker/Zenk-Möltgen/Mauer 2017: 137).

9.4 Software zur Erfassung und Bearbeitung von Metadaten

Für die Verwaltung von Metadaten in einer komplexen Struktur wie DDI-Lifecycle ist die Unterstützung von Software erforderlich, da eine manuelle Codierung der XML-Strukturen unter Verwendung sehr vieler Referenzen aufwändig ist. Im Folgenden werden daher einige Programme kurz vorgestellt, die ein Management der Metadaten erlauben und die für Forschende in den Sozialwissenschaften eine relevante Rolle spielen. Mit Hilfe dieser Programme lassen sich die Metadaten nicht nur erstellen und editieren, sondern auch publizieren und durchsuchen. Die Auswahl berücksichtigt dabei besonders kostenfrei verwendbare Programme und erhebt keinen Anspruch auf Vollständigkeit. Weitere Software zur Nutzung des DDI Standard ist auf den Webseiten der DDI Alliance aufgeführt.

9.4.1 Colectica for Excel

Für die direkte Bearbeitung von Metadaten auf Studien- und Variablenebene in Microsoft Excel gibt es die Erweiterung *Colectica for Excel*. In der kostenfreien Standardversion lassen sich Datensätze mit DDI-Lifecycle-kompatiblen Beschreibungen versehen, die jedoch nur ganz grundlegende Elemente enthalten. Ebenso können Wertelisten, die etwa Antwortwerte zu Fragen enthalten, wiederverwendet und dokumentiert werden. In der kostenpflichtigen Professional Version können Daten und Metadaten aus SPSS, STATA und SAS nach Excel importiert werden. Der Vorteil ist, dass die Excel-Datei bei der Weitergabe dann auch immer die Metadaten enthält und so die Daten für Dritte besser verständlich sind. Weitere kostenpflichtige Produkte von Colectica ermöglichen die Bearbeitung komplexer DDI-Lifecycle-Dokumentationen.

9.4.2 Dataset Documentation Manager (DSDM)

Das Programm *Dataset Documentation Manager* (DSDM) ist speziell für die Beschreibung von Datensätzen geeignet (Zenk-Möltgen 2006) und mit dem DDI-Codebook Standard kompatibel. Es handelt sich um ein Stand-Alone Windows-Programm und kann SPSS-Datensätze einlesen. Anschließend kann die Dokumentation der Variablen, Fragetexten, Antwortvorgaben, Codes, Interviewerанweisungen etc. erfolgen. DSDM bietet eine Unterstützung von Mehrsprachigkeit und Unicode. Ein Export der Metadaten in verschiedene DDI-Formate ist möglich. Zusätzlich kann mit dem Programm eine Erstellung von *CodebookExplorer*

Datenbanken durchgeführt werden – diese ermöglichen eine Suche in Metadaten, einen Vergleich und die Kategorisierung von Variablen, sowie eine einfache Analyse der Daten.

9.4.3 DBKfree und DBKForm

Die Software des Datenbestandskatalogs von GESIS wird als *DBKfree* in einer Open-Source-Variante zur Verfügung gestellt. Sie ermöglicht die Erfassung von Metadaten für Studienbeschreibungen durch die Komponente *DBKEdit* bzw. für die Recherche im Web durch die Komponente *DBKSearch*. Das einfache Datenmodell des DBK ist kompatibel zu DDI-Codebook, die Metadaten werden als Export jedoch auch in DDI-Lifecycle zur Verfügung gestellt. Für Nutzer bietet *DBKfree* eine einfache und eine fortgeschrittene Suche, das Blättern in den Studienbeschreibungen und eine Anzeige nach verschiedenen Kriterien wie Themenkategorien, Primärforschenden oder geographischen Einheiten.

Ein sehr einfaches freies Tool für die Erfassung von Metadaten auf Studienebene ist *DBKForm*. Es besteht aus einem HTML-Formular, welches lokal aufrufbar ist und eine DDI-Codebook kompatible XML-Datei erzeugt. Diese kann bei der Verwendung von *DBKfree* für einen Import oder für eine anderweitige Verarbeitung der DDI-Codebook XML-Datei genutzt werden. DBKForm unterstützt die grundlegenden Metadaten für eine Studienbeschreibung und unterstützt die Verwendung eines kontrollierten Vokabulars für einige der Elemente.

9.4.4 Nesstar Publisher und Server

Ein komfortabler Editor für DDI-Codebook-kompatible Metadaten ist der kostenpflichtige *Nesstar Publisher*. Über Eingabeformulare können umfangreiche Metadaten zur Studie und zu den Variablen des Datensatzes eingegeben werden. Diese werden als XML-Datei im DDI-Codebook Format abgespeichert und können auch für die Publikation der Studie auf einem separat erhältlichen *Nesstar Server* genutzt werden. Der *Nesstar Server* bietet neben einer Weboberfläche für die Anzeige und Recherche in den Metadaten auch die Möglichkeit einer Onlineanalyse. Zusätzlich verfügbar sind verschiedene Programmierschnittstellen (*Application Programming Interface*, API) zur Anbindung an weitere Systeme (etwa eine Public API, eine REST API oder einen OAI-PMH Server).

9.4.5 Dataverse

Ein komplett als Open Source verfügbares System zur Dokumentation, Präsentation von und Suche in Forschungsdaten ist das *Dataverse Project*. Für die Entwicklung haben sich das Institute for Quantitative Social Science (IQSS), die Harvard University Library und die Harvard University Information Technology zusammengeschlossen. Institutionen können mit der Software eine eigene Instanz eines Dataverse-Servers erstellen, in welchem die Forschenden ihre Daten beschreiben, sichern und teilen können. Wissenschaftler/innen haben auch die Möglichkeit, nach Registrierung das *Harvard Dataverse* zu verwenden, ohne eine eigene Softwareinstanz installieren zu müssen. Eine Beschreibung von Daten in Dataverse ermöglicht die Zitation und eine ausführliche Beschreibung auf Studienebene, die zu Standards wie etwa DDI-Codebook, DataCite und Dublin Core kompatibel sind.

9.4.6 Archivist

Ein weiteres Open-Source-Tool für die Dokumentation von Fragebögen ist die Software *Archivist*, die von CLOSER (Cohort & Longitudinal Studies Enhancement Resources) entwickelt wurde. Es ermöglicht die Dokumentation von Fragebögen im DDI-Lifecycle Standard ebenso wie den Austausch entsprechender Metadaten mit anderen Programmen, sodass die Software in einen Workflow für die Datendokumentation integriert werden kann.

9.4.7 Ced²ar

Für die Publikation und Suche im Web ist die Open-Source-Software *Ced²ar* der Cornell University entwickelt worden. Ced²ar steht für *Comprehensive Extensible Data Documentation and Access Repository* und basiert auf dem DDI-Codebook Standard. Sie bietet neben der Anzeige von Metadaten auf Studienebene auch eine Suche auf der Variablenebene und einen Vergleich von Variablenbeschreibungen.

9.5 Fazit

Die Darstellung von Metadatenstandards und ihrer Anwendungsmöglichkeiten im Kontext sozialwissenschaftlicher Forschungsdaten zeigt, dass eine recht große Bandbreite zwischen sehr einfacher und sehr ausführlicher Dokumentation auf Studien- und Variablenebene vorhanden ist. Für Forschende ergibt sich daraus die Herausforderung, die für die eigenen Anwendungsanforderungen geeignete Mischung von Standard und Software auszuwählen. In diesem Zusammenhang sollte die Forderung, Replikationen von Forschungsergebnissen zu ermöglichen, dazu führen, dass ein Mindestmaß an Dokumentation für diesen Zweck bereitgestellt wird. Ein Anlass zu einer erweiterten Dokumentation kann sein, dass auch anderen Forschenden die Möglichkeit eröffnet werden soll, weitere neue Forschungsfragen anhand eines erhobenen Datensatzes zu beantworten. Dass auch zunehmend die Erhebung, das Management und die Erstellung von Dokumentationen zu Forschungsdaten in der wissenschaftlichen Community als eigenständige und zitierfähige Forschungsleistung anerkannt werden, ist sicher ein zusätzlicher Anreiz, an einer qualitativ hochwertigen Dokumentation von Forschungsdaten zu arbeiten.

Literaturverzeichnis

Akdeniz, Esra/Zenk-Möltgen, Wolfgang (2017): DDI-Lifecycle im Datenarchiv. Das Metadatenschema für die Dokumentation in verschiedenen Softwaresystemen. GESIS Papers 2017/02. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-50354-9 [Zugriff: 20.06.2018].

Bauske, Franz (2000): Das Studienbeschreibungsschema des Zentralarchivs. In: ZA-Information 47, S. 73-80. https://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/za_information/ZA-Info-47.pdf [Zugriff: 20.06.2018].

Bersheim, Sabrina/Oschmiansky, Frank/Sell Stefan (2014): Wie wird Arbeitslosigkeit gemessen? https://www.bpb.de/politik/innenpolitik/arbeitsmarktpolitik/54909/arbeitslosigkeit-messen?p=all [Zugriff: 20.06.2018].

Corti, Louise/ Van den Eynden, Veerle /Bishop, Libby/Woollard, Matthew (2014): Managing and Sharing Research Data. A Guide to Good Practice. London: Sage Publications.

DataCite (2016): DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. DataCite e.V. http://doi.org/10.5438/0012.

DCMI (2011): Resource Discovery, Resource Description. Archived MediaWiki Page. https://github.com/dcmi/repository/blob/master/mediawiki_wiki/Glossary/Resource_Discovery.md [Zugriff: 20.06.2018].

DDI (o.J.): Lifecycle-Modell: http://www.ddialliance.org/sites/default/files/what-is-ddi-diagram.jpg [Zugriff: 20. 06.2018].

Ebel, Thomas (2016): Einreichung von Syntaxen in datorium (Replikationsserver). http://www.replikationsserver.de/replikationsserver/home/publikationen [Zugriff: 20.06.2018].

Ebel, Thomas/Trixa, Jessica (2015): Hinweise zur Aufbereitung quantitativer Daten. GESIS Papers 2015/09. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-432235 [Zugriff: 20.06.2018].

Edwards, Paul N./Mayernik, Matthew S./Batcheller, Archer L./Bowker, Geoffey C./Borgman, Christine L. (2011): Science Friction. Data, Metadata, and Collaboration. In: Social Studies of Science 41, 5, S. 667-690.

Gartner, Richard (2008): Metadata for Digital Libraries. State of the Art and Future Directions. JISC Technology and Standards Watch Reports. http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf [Zugriff: 20.06.2018].

GESIS – Leibniz-Institut für Sozialwissenschaften (2018): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016. GESIS Datenarchiv, Köln. ZA5250 Datenfile Version 2.1.0, https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5250&db=d&doi=10.4232/1.12796 [Zugriff: 20.06.2018].

Gilliland, Anne J. (2016): Setting the Stage. In: Baca, Murtha (Hrsg.): Introduction to Metadata. Los Angeles: Getty Publications, 2016. https://www.getty.edu/publications/intrometadata/setting-the-stage/ [Zugriff: 20.06.2018].

Hausstein, Brigitte/Zenk-Möltgen, Wolfgang (2011): da|ra. Ein Service der GESIS für die Zitation sozialwissenschaftlicher Daten. In: Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung vom 20./21. September 2010, S. 139-147. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf [Zugriff: 20.06.2018].

Helbig, Kerstin/Hausstein, Brigitte/Toepfer, Ralf (2015): Supporting Data Citation. Experiences and Best Practices of a DOI Allocation Agency for Social Sciences. In: Journal of Librarianship and Scholarly Communication 3(2), p.eP1220. https://doi.org/10.7710/2162-3309.1220. [Zugriff: 20.06.2018].

Horstmann, Wolfram (2007): Open Access international. Lokale Systeme, kooperative Netzwerke und visionäre Infrastrukturen. In: Zeitschrift für Bibliothekswesen und Bibliographie 54, 4/5, S. 230-233. https://edoc.hu-berlin.de/bitstream/handle/18452/10000/18.pdf?sequence=1 [Zugriff: 20.06.2018].

Hughes, Keith/Jenkins, Stephen/Wright, Geoff (2000): triple-s XML. A Standard Within a Standard. In: Social Science Computer Review 18, 4, S. 421-433.

Jääskeläinen, Taina/Moschner, Meinhard/Wackerow, Joachim (2009): Controlled Vocabularies for DDI 3. Enhancing Machine-Actionability. In: IASSIST Quarterly Spring/Summer 2009, S. 34-39. http://www.iassistdata.org/sites/default/files/iqvol3312wackerow_0.pdf [Zugriff: 20.06.2018].

Jensen, Uwe/Katsanidou, Alexia/Zenk-Möltgen, Wolfgang (2011): Metadaten und Standards. In: Büttner Stephan/Hobohm, Hans-Christoph/Müller L. (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock u. Herchen, S. 83-100. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:kobv:525-opus-2318 [Zugriff: 20.06. 2018].

Jensen, Uwe (2012): Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten. GESIS-Technical Reports 2012/07. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-320650 [Zugriff: 20.06. 2018].

Jensen, Uwe/Ekman, Stefan/Hjelm, Claus-Göran/Irebäck, Hans/Schweers, Stefan (2015): DwB (Data without Boundaries) Report D7.6. Metadata Standards and Practices in Related Disciplines and Standards for Linking Different Sources. http://doi.org/10.13140/RG.2.1.4164.8248.

Karjalainen, Merja/Kleemola, Mari/Jensen, Uwe (2012): DwB (Data without Boundaries) Report D7.1: Metadata standards. Usage and Needs in NSIs and Data Archives. http://doi.org/10.13140/RG.2.1.2198.7442.

Koch, Ute/Akdeniz, Esra/Meichsner, Jana/Hausstein, Brigitte/Harzenetter, Karoline (2017): da|ra Metadata Schema. Documentation for the Publication and Citation of Social and Economic Data. Version 4.0. GESIS Papers 2017/25. http://dx.doi.org/10.4232/10.mdsdoc.4.0.

Lebo, Timothy/Gregory, Todd W. (2010): Converting Governmental Datasets into Linked Data. In: Proceedings of the 6th International Conference on Semantic Systems. New York: ACM Digital Library. https://doi.org/10.1145/1839707.1839755.

Netscher, Sebastian/Eder, Christina (Hrsg.) (2018): Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research. GESIS Papers, 2018/22. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59492-3 [Zugriff 26.10.2018].

Neuroth, Heike/Strathmann, Stefan/Oßwald, Achim/Scheffel, Regine/Klump, Jens/Ludwig Jens (Hrsg.) (2012): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. http://nbn-resolving.de/urn:nbn:de:0008-2012031401 [Zugriff: 20.06.2018].

Nurseitov, Nurzhan/Paulson, Michael/Reynolds, Randall/Izurieta, Clemente (2009): Comparison of JSON and XML Data Interchange Formats. A Case Study. In: Proceedings of the ISCA 22nd International Conference on Computer Applications in Industry and Engineering, CAINE 2009, S. 157-162.

Recker, Jonas/Zenk-Möltgen, Wolfgang/Mauer, Reiner (2017). Applications of Research Data Management at GESIS Data Archive for the Social Sciences. In: Kruse, Fillip /Thestrup, Jesper Boserup (Hrsg.): Research Data Management - A European Perspective. Berlin, Boston: De Gruyter, S. 119 - 146. https://doi.org/10.1515/9783110365634-008.

Riley, Jenn (2017): Understanding Metadata. What is Metadata, and what is it for? Washington DC: National Information Standards Organization. http://www.niso.org/publications/understanding-metadata-2017 [Zugriff: 20.06.2018].

Rühle, Stefanie (o.J.): Kleines Handbuch Metadaten. Metadaten. http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Handbuch/metadaten.pdf?__blob=publicationFile [Zugriff: 20.06.2018].

Siegl, Johannes (o.J.): Online Umfrage Software im Vergleich. https://trusted.de/online-umfrage [Zugriff: 28.10.2018].

Starr, Joan/Gastl, Angela (2011): isCitedBy. A Metadata Scheme for DataCite. In: D-Lib Magazine 17, 1/2. http://www.dlib.org/dlib/january11/starr/01starr.html [Zugriff: 20.06.2018].

Watteler, Oliver (2010): Erstellung von Methodenberichten für die Archivierung von Forschungsdaten. http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Aufbau_Methodenbericht_v1_2010-07.pdf [Zugriff: 20.06.2018].

Zenk-Möltgen, Wolfgang (2006): Dokumentation von Umfragedaten in Länder vergleichender Perspektive mit Hilfe des ZA Dataset Documentation Managers (DSDM). In: ZA-Information/Zentralarchiv für Empirische Sozialforschung 59, S. 159-170. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-198427 [Zugriff: 20.06.2018].

Zenk-Möltgen, Wolfgang (2012): Metadaten und die Data Documentation Initiative (DDI). In: Altenhöner, Reinhard/Oellers, Claudia (Hrsg.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Berlin: Scivero, S. 111-126. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-46679-8 [Zugriff: 20.06.2018].

Zenk-Möltgen, Wolfgang/Habbel, Norma (2012): Der GESIS Datenbestandskatalog – und sein Metadatenschema. Version 1.8. GESIS-Technical Reports 2012/01. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-292372 [Zugriff: 20.06.2018].

Linkverzeichnis

EUDAT: B2Find Katalog: http://b2find.eudat.eu/ [Zugriff: 20.06.2018].

CESSDA – Council of European Social Science Data Archives: https://cessda.net/ [Zugriff: 20.06.2018].

CLOSER – Cohort & Longitudinal Studies Enhancement Resources: https://www.closer.ac.uk/ [Zugriff: 20.06.2018].

CLOSER Software Archivist: https://github.com/CLOSER-Cohorts/archivist [Zugriff: 20.06.2018].

CodebookExplorer: https://dbk.gesis.org/software/cbe.asp [Zugriff: 20.06.2018].

Colectica für Excel: http://www.colectica.com/software/colecticaforexcel [Zugriff: 20.06.2018].

Cornell University Software Ced2ar – Comprehensive Extensible Data Documentation and Access Repository: http://www2.ncrn.cornell.edu/ced2ar-web/about [Zugriff: 20.06.2018].

CSV-Format: https://de.wikipedia.org/wiki/CSV_%28Dateiformat%29 [Zugriff: 20.06.2018].

da|ra OAI-PMH: http://www.da-ra.de/oaip/oai?verb=ListMetadataFormats [Zugriff: 20.06.2018].

DARIAH-DE: https://wiki.de.dariah.eu/display/publicde/5.+Kontrolliert-Strukturierte+Vokabulare [Zugriff: 20.06.2018].

Data Documentation Initiative (DDI): http://www.ddialliance.org [Zugriff: 20.06.2018].

DataCite: https://www.datacite.org/ [Zugriff: 20.06.2018].

DataCite Metadatenschema 4.1 (DataCite 2017): https://schema.datacite.org/meta/kernel-4.1/ [Zugriff: 20.06.2018].

Dataset Documentation Manager (DSDM): https://dbk.gesis.org/software/dsdm.asp [Zugriff: 20.06.2018].
Dataverse Appendix: http://guides.dataverse.org/en/latest/user/appendix.html [Zugriff: 20.06.2018].
Dataverse Project: https://dataverse.org/about/ [Zugriff: 20.06.2018].
DBKEdit: https://dbk.gesis.org/dbkform/ [Zugriff: 20.06.2018].
DBKfree: https://dbk.gesis.org/DBKfree2.0/ [Zugriff: 20.06.2018].
DCMI: Type Vocabulary (2012): http://dublincore.org/documents/dcmi-type-vocabulary/ [Zugriff: 20.06.2018].
DDI: CV: https://www.ddialliance.org/controlled-vocabularies [Zugriff: 20.06.2018].
DDI: Publikationen: http://www.ddialliance.org/resources/publications [Zugriff: 20.06.2018].
DDI: Software: http://www.ddialliance.org/resources/tools [Zugriff: 20.06.2018].
DDI: User Conference (EDDI): http://www.eddi-conferences.eu [Zugriff: 20.06.2018].
DDI: Workshops: https://www.ddialliance.org/training [Zugriff: 20.06.2018].
DDI-C Standard: http://www.ddialliance.org/Specification/DDI-Codebook/2.5 [Zugriff: 20.06.2018].
DDI-L Standard: http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2 [Zugriff: 20.06.2018].
DOI –Digital Object Identifier: https://www.doi.org/ [Zugriff: 20.06.2018].
DCMES – Dublin Core Metadata Element Set (2012): ISO 15836: http://dublincore.org/documents/dces/ [Zugriff: 20.06.2018].
DCMI – Dublin Core Metadata Initiative: http://dublincore.org [Zugriff: 20.06.2018].
Export SPSS to Triple-S: http://spsstools.net/en/scripts/830/ [Zugriff: 20.06.2018].
GND – Gemeinsame Normdatei: http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html [Zugriff: 20.06.2018].
GESIS DBK – Datenbestandskatalog: https://dbk.gesis.org/dbksearch/home.asp?db=d [Zugriff: 20.06.2018].
GESIS datorium: https://datorium.gesis.org/ [Zugriff: 20.06.2018].
Harvard Dataverse: https://dataverse.harvard.edu/ [Zugriff: 20.06.2018].
IdSurvey: http://www.idsurvey.com/analyze/ [Zugriff: 20.06.2018].
Inter-university Consortium for Political and Social Research (ICPSR): http://www.icpsr.umich.edu [Zugriff: 20.06.2018].
Bitte ersetzen durch:
ISCED (2011): International Standard Classification of Education. UNESCO Institute for Statistics: http://uis.unesco.org/en/topic/international-standard-classification-education-isced [Zugriff: 20.06.2018].
ISCO – International Standard Classification of Occupation: http://www.ilo.org/public/english/bureau/stat/isco/ [Zugriff: 20.06.2018].
ISO 15836 Information and Documentation. The Dublin Core Metadata Element Set, Part 1: Core Elements: https://www.iso.org/standard/71339.html [Zugriff: 20.06.2018].
ISO 3166 Country Codes: https://www.iso.org/iso-3166-country-codes.html [Zugriff: 20.06.2018].
ISO 639 Language codes at ISO: https://www.iso.org/iso-639-language-codes.html [Zugriff: 20.06.2018].
ISO 639 Language codes at Library of Congress: https://www.loc.gov/standards/iso639-2/php/code_list.php [Zugriff: 20.06.2018].
JSON – JavaScript Object Notation: https://de.wikipedia.org/wiki/JavaScript_Object_Notation [Zugriff: 20.06.2018].
LimeSurvey: https://www.limesurvey.org/community/extensions/97-export-to-triple-s-survey-interchange-standard [Zugriff: 20.06.2018].
Nesstar Publisher: http://www.nesstar.com/software/publisher.html [Zugriff: 20.06.2018].
Nesstar Server: http://www.nesstar.com/software/server.html [Zugriff: 20.06.2018].
NIPO Online, CAPI and CATI Survey Software Solutions: https://www.nipo.com/ [Zugriff: 20.06.2018].
OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting: https://www.openarchives.org/pmh/ [Zugriff: 20.06.2018].
OAI-PMH Dublin Core (oai_dc): https://www.openarchives.org/OAI/openarchivesprotocol.html#dublincore [Zugriff: 20.06.2018].
Online-Umfrageprogramme at websm.org 2018: http://www.websm.org/c/1283/Software/ [Zugriff: 20.06.2018].
OpenAire: https://www.openaire.eu/search/find?keyword [Zugriff: 20.06.2018].
R Paket sss. Tools for Importing Files in the Triple-s Format: https://cran.r-project.org/web/packages/sss/sss.pdf [Zugriff: 20.06.2018].
Registrierungsagentur da|ra: https://www.da-ra.de [Zugriff: 20.06.2018].
RFC3629. Spezifikation des Unicode-Zeichenformat UTF-8: https://tools.ietf.org/html/rfc3629 [Zugriff: 20.06.2018].
RFC4180. Spezifikation des Dateiformates CSV: http://tools.ietf.org/html/rfc4180 [Zugriff: 20.06.2018].
SensusWeb: http://www.sawtooth.com/index.php/software/sensus-web/evalsys/ [Zugriff: 20.06.2018].

SnapSurveys: https://www.snapsurveys.com/support/data-management/ [Zugriff: 20.06.2018].
STW – Standard-Thesaurus Wirtschaft: http://zbw.eu/stw/version/latest/ [Zugriff: 20.06.2018].
Statistik-Software: R, SAS, SPSS und STATA im Vergleich (2018): https://www.inwt-statistics.de/blog-artikel-lesen/Statistik-Software-R_SAS_SPSS_STATA_im_Vergleich.html [Zugriff: 20.06.2018].
Thesaurus Sozialwissenschaft (TheSoz): http://lod.gesis.org/thesoz/de/hierarchical_concepts.html [Zugriff: 20.06.2018].
Triple-S – the Survey Interchange Standard: http://www.triple-s.org/ [Zugriff: 20.06.2018].
Triple-S 2017: http://www.triple-s.org/wp-content/uploads/Triple-S-XML-3.0-Release-Notes.pdf [Zugriff: 20.06.2018].
Trost Media 2018. Zeilenumbruch: https://www.sttmedia.de/zeilenumbruch [Zugriff: 20.06.2018].
XML – Extensible Markup Language: https://de.wikipedia.org/wiki/Extensible_Markup_Language [Zugriff: 20.06.2018].